

Article

Excavator Activity Recognition Using Gramian Angular Field (GAF) Encoding of Multi-Sensor Operational Data and Deep Learning

Abubakar Sharafat ^{1,*} , Abid Ullah ¹, Waqas Arshad Tanoli ^{2,*}  and Saad Arif ³ 

¹ Department of Civil and Environmental Engineering, Hanyang University, Seoul 04763, Republic of Korea; abid437@hanyang.ac.kr

² Department of Civil & Environmental Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia

³ Department of Mechanical Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia; sarif@kfu.edu.sa

* Correspondence: abubakarsharafat89@gmail.com (A.S.); wtanoli@kfu.edu.sa (W.A.T.)

Abstract

Efficient monitoring and recognition of excavator operational activities are critical for improving productivity, safety, and intelligent construction management. Traditional vision-based activity recognition systems remain sensitive to challenging construction site conditions, including occlusion, dust, variable illumination, and limited visibility. This paper presents an excavator activity recognition framework that, for the first time, applies Gramian Angular Field (GAF) encoding to multi-sensor excavator operational data, transforming them into spatial image representations and enabling deep convolutional neural networks (CNNs) to extract discriminative features. The proposed approach integrates synchronized excavator sensor signals including bucket positional coordinates, body orientation, fuel consumption, engine RPM, and joint angles to characterize operational behavior during four representative activities: digging, dumping, idle, and levelling. Unlike conventional sensor-based methods that directly process sequential time-series data, our GAF-based framework transforms operational signals into structured spatial representations that preserve temporal correlations while enabling effective feature learning through image-based deep learning architectures. Experimental results demonstrate that the proposed GAF-CNN-LSTM framework achieves 8.24 percentage points higher classification accuracy compared with LSTM networks, and 6.24 percentage points higher than 1D CNN baselines trained on the same sensor data. The method effectively captures discriminative operational signatures across multiple excavator activities in the collected dataset. This work bridges the limitations of both vision-based and traditional sensor-based approaches, providing a promising framework for excavator activity monitoring in construction and fleet management contexts, pending further validation under operational deployment conditions.



Academic Editor: Ke Chen

Received: 1 September 2026

Revised: 25 September 2026

Accepted: 25 September 2026

Published: 27 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: excavator activity recognition; Gramian angular field; time-series imaging; CNN-LSTM; multi-sensor data; earthwork operations; smart construction

1. Introduction

Earthwork operations are among the most critical and resource-intensive activities in construction projects, where excavators play a dominant role in excavation, loading,

dumping, and grading tasks [1,2]. These operations are further complicated by heterogeneous soil and ground conditions, whose dynamic mechanical behavior has been widely investigated in geotechnical and construction materials research [3–6]. Efficient monitoring and recognition of excavator operational activities are essential for improving productivity, safety, equipment utilization, and intelligent construction management [7]. In recent years, the rapid advancement of construction automation and smart construction technologies has increased the demand for reliable activity recognition systems capable of understanding equipment behavior in real time [8]. Comparable data-driven monitoring and automated inspection frameworks have already been demonstrated across civil and infrastructure engineering, including crack detection using unmanned devices [9], automatic segmentation of bridge point clouds [10], robotic inspection of underground utilities and large-scale freeform parts [11,12], and real-time structural response prediction of offshore bridges [13]. Accurate identification of excavator activities can support autonomous construction equipment, fleet management, productivity analysis, digital twin systems, and equipment safety monitoring.

Traditional excavator activity recognition approaches have primarily relied on vision-based monitoring systems using computer vision and deep learning techniques. With the development of convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) models, and CNN-LSTM hybrid architectures, significant progress has been achieved in recognizing construction equipment activities from video streams [14]. Vision-based methods provide rich spatial information and have demonstrated strong performance in activity classification tasks [15]. However, these approaches remain highly sensitive to practical construction site conditions, including occlusion, dust, unstable illumination, weather variability, camera viewpoint dependency, and limited visibility [16]. In earthwork environments, excavator operations are frequently partially or fully obstructed by surrounding equipment, soil piles, workers, or terrain conditions, significantly degrading the robustness and reliability of camera-based recognition systems [1]. Although advanced image enhancement techniques can partially alleviate degraded imaging conditions such as low illumination [17], they cannot fully resolve severe occlusion and viewpoint limitations.

To address the limitations of vision-based approaches, sensor-based activity recognition methods have gained increasing attention in recent years. Various operational sensors, including inertial measurement units (IMUs), accelerometers, gyroscopes, hydraulic pressure sensors, engine signals, and kinematic measurements, have been utilized to monitor excavator operational behavior [18–20]. Compared with camera-based systems, sensor-based methods are less affected by environmental visibility conditions and can provide stable operational measurements during construction activities. Furthermore, modern construction equipment is increasingly equipped with onboard sensors and telematics systems, making sensor-based monitoring more practical for real-world deployment [20–22]. Continuing advances in sensing technologies including flexible strain sensors [23], biomimetic tactile sensing systems [24], and precisely calibrated magnetic sensor arrays [25] are further expanding the range and fidelity of measurable operational signals. Nevertheless, existing sensor-based activity recognition methods often rely on handcrafted feature extraction or direct sequential processing of time-series data using conventional machine learning or recurrent neural network architectures. These approaches may have limited capability to effectively capture complex temporal correlations and discriminative operational patterns embedded within multi-sensor excavator signals.

Recently, time-series imaging techniques have emerged as a promising approach for converting one-dimensional sequential signals into two-dimensional image representations suitable for deep learning-based feature extraction. Among these techniques, GAF

transformation has shown considerable potential [26] for preserving temporal dependency and correlation structures within time-series data while enabling convolutional neural networks to learn highly discriminative spatial representations. By transforming sequential sensor signals into image-like representations, GAF encoding enables deep learning models to extract texture, shape, and temporal correlation patterns that are difficult to capture using conventional sequential processing methods. Although GAF-based learning has demonstrated promising results in several engineering and industrial applications [27,28], its application to excavator activity recognition in earthwork operations remains largely unexplored. The growing adoption of time-series methods across construction management domains—spanning cost forecasting, safety monitoring, and productivity assessment—reflects a broader trend toward data-driven intelligence in the built environment [29].

Taken together, these three research directions point to different gaps. Vision-based methods are limited mainly by unreliable sensing, i.e., occlusion, dust, and lighting. Sensor-based methods are limited because they rely on simple, handcrafted features taken from raw signal sequences. Time-series imaging methods such as GAF solve the representation problem directly but have not yet been tested on excavator sensor data. This study therefore targets a data-representation gap rather than an architectural one, and the boundary of the claim is stated explicitly. The Gramian Angular Field transform, the pairing of GAF images with convolutional backbones, and hybrid CNN-LSTM classifiers are all established methods, and no novelty is claimed for any of them individually. What is specific to this work is threefold. First, ten heterogeneous excavator channels spanning positional, orientational, and powertrain measurements are concatenated into a single Gramian Angular Difference Field (GADF), so that off-diagonal blocks expose cross-channel angular relationships to the convolution directly, rather than encoding one channel per image or one modality per stream as in prior single-modality GAF studies. Second, the encoded field is treated as a temporal sequence, so that the recurrent stage models transitions between successive encoded windows rather than classifying a single static image. Third, the contribution of the encoding is examined experimentally by evaluating a CNN-LSTM on raw sequences and a Transformer encoder on both raw and GADF inputs; because neither pair of models was trained under fully matched configurations, these comparisons indicate, rather than isolate, the contribution of the representation. The proposed framework utilizes synchronized excavator sensor signals, including bucket positional coordinates (X, Y, and Z), body pitch, body roll, fuel consumption, engine RPM, arm slope, boom slope, and bucket slope, to characterize excavator operational behavior during earthwork activities. The collected time-series sensor data are preprocessed and transformed into GADF representations, which are subsequently utilized as inputs for CNN-based activity classification. Four representative excavator activities, including digging, dumping, idle, and levelling, are investigated in this study. Unlike conventional sensor-based methods that directly process raw sequential signals, the proposed framework transforms operational sensor data into structured spatial representations, enabling more effective spatial-temporal feature learning through deep convolutional architectures.

The proposed approach aims to overcome the limitations of both vision-based and traditional sensor-based activity methods by combining the visibility-independence of onboard sensor monitoring with the representation-learning capability of image-based deep learning. Experimental results demonstrate that the proposed GAF-based framework significantly improves excavator activity recognition performance, achieving 6.24–8.24 percentage points higher classification accuracy than the raw time-series baselines evaluated in this study. The results further indicate that the proposed method effectively captures discriminative operational signatures across multiple excavator activities in the collected dataset.

The main contributions of this study are summarized as follows:

- A joint multi-channel GADF encoding is proposed, in which ten heterogeneous excavator channels are concatenated into a single Gramian Angular Difference Field so that cross-channel angular relationships appear as off-diagonal structure. This encoding strategy, not the GAF transform or the CNN-LSTM classifier themselves, is the element claimed as new.
- Excavator time-series sensor signals are converted into image-like spatial representations to enable deep convolutional feature extraction.
- Multiple operational, kinematic, and orientation-related sensor signals are integrated for activity classification.
- The proposed framework is comprehensively compared with LSTM, 1D CNN, and Transformer baselines operating on raw time-series data, and with a Transformer applied to the same GADF image sequences.
- The proposed GAF-based deep learning framework achieves 6.24–8.24 percentage points higher classification accuracy than the raw time-series baselines evaluated in this study.

2. Related Work

Vision-based activity recognition has been the most widely studied paradigm for monitoring construction equipment [30]. Early studies relied on hand-engineered visual features and classical classifiers to detect equipment poses and actions from site imagery, whereas more recent work has adopted deep architectures—convolutional neural networks for spatial feature extraction and recurrent or two-stream networks for temporal modelling—to classify earthmoving activities directly from video [7,16,31–35]. These methods can localize multiple machines and recognize fine-grained actions when an unobstructed view is available. In practical earthwork settings, however, their accuracy degrades under occlusion by soil piles, dust, and co-working equipment, as well as under variable illumination and camera viewpoints, and they additionally raise installation, calibration, and privacy concerns on active job sites [36]. The reliability of such pipelines is further bounded by the quality of the captured visual data, which has motivated dedicated research on video restoration and image deblurring [37,38] and on quality assessment of 3D point-cloud data [39,40].

To overcome these visibility constraints, a growing body of research has investigated sensor-based recognition using signals acquired directly from the machine. Inertial measurement units, accelerometers, and gyroscopes mounted on the boom, arm, or cab have been used to infer operational states [20], and hydraulic-pressure, engine, and kinematic channels have been exploited to characterise the work cycle [20–22,41–45]. Song et al. [46] proposed a hybrid multi-scale feature extractor combining CNN temporal self-attention and 1D-ResNet50 with decision-level data fusion for excavator working cycle stage identification using hydraulic pressure signals, achieving 95.21% accuracy. Reported approaches typically extract statistical or frequency-domain features from sliding windows and classify them with support vector machines, random forests, or hidden Markov models, or process the raw sequences with recurrent and long short-term memory networks [47]. Although these techniques are robust to environmental visibility, handcrafted features depend heavily on expert design, and purely sequential models can struggle to capture the higher-order correlations that distinguish operationally similar activities such as dumping and levelling.

A complementary line of research encodes one-dimensional time series as two-dimensional images so that the representational power of convolutional networks can be brought to bear on sequential data. Recurrence plots, Markov transition fields, short-time Fourier and wavelet spectrograms, and Gramian Angular Fields have all been

used to transform signals into images for fault diagnosis, human activity recognition, and biomedical classification [25,48]. Among these, the Gramian Angular Field is attractive because it encodes the temporal correlation structure of a signal in a polar-coordinate Gram matrix, producing images whose texture reflects the trajectory and dynamics of the underlying series while remaining information-preserving [49]. Such encodings have achieved strong results across several engineering domains, yet their use for construction-equipment activity recognition from multi-channel operational data has received little attention. Where GAF encoding has been applied in the construction domain, it has so far been used on single-modality inertial signals from body-worn or machine-mounted units, and the question of how heterogeneous operational channels should be jointly encoded within a single field remains open.

Beyond the choice of sensing modality, the motivation for excavator activity recognition is closely tied to automated productivity analysis and project control. Accurate, continuous identification of work states enables cycle-time estimation, idle and downtime detection, fuel- and emission accounting, and automated progress reporting that would otherwise rely on manual observation or self-reported logs [22]. Earlier construction-management studies estimated equipment productivity through time-and-motion studies or radio-frequency and satellite-positioning tracking, but these provide only coarse location or utilisation information and cannot distinguish the fine-grained actions that constitute an earthmoving cycle [50,51]. This practical demand for granular, machine-level activity data has driven the shift from manual monitoring toward automated recognition methods, and it frames activity classification as a foundational capability for downstream analytics rather than an end in itself. Beyond productivity, time-series forecasting methods have also been applied to construction safety management, where hybrid decomposition and deep learning frameworks have been used to forecast fatal accident trends and support proactive risk mitigation [52].

Researchers have also explored sensing modalities beyond vision and motion. Acoustic and vibration signals generated by the engine, hydraulics, and tool-ground interaction have been used to infer operating states, exploiting the distinct sound and vibration signatures of digging, travelling, and idling [22,41]; such signatures ultimately stem from the dynamic behaviour of the machine's mechanical and transmission components [53]. Satellite-positioning and telematics streams, now standard on modern fleets, capture machine location, travel, and load cycles at low cost, although their temporal resolution limits fine-action discrimination [50]. To combine the complementary strengths of these sources, multi-sensor and multi-modal fusion strategies have been proposed, integrating inertial, kinematic, acoustic, and positioning channels at the feature or decision level [54–56]. Fusion generally improves robustness, but it also increases model complexity and the burden of temporal synchronisation, and many reported systems still depend on handcrafted features or shallow classifiers that do not fully exploit the joint structure of the combined signals.

Within the time-series-imaging paradigm, the design of the downstream classifier has received considerable attention. Because encoded signals resemble natural images, deep convolutional networks pretrained on large image corpora can be transferred and fine-tuned, allowing strong performance even when labelled operational data are scarce [57]. Residual networks and other deep backbones have been applied to spectrogram and Gramian Angular Field inputs for machinery fault diagnosis and human activity recognition, frequently surpassing models trained directly on the raw sequences [58,59]. Studies have further compared the summation and difference variants of the Gramian Angular Field (GASF and GADF) and combined them with Markov transition fields as multi-channel inputs, while more recent work has introduced attention mechanisms and vision Transformers to capture long-range dependencies within the encoded images [60]. These findings

collectively indicate that pairing an information-preserving encoding with an expressive convolutional backbone is a particularly effective recipe for sequential-signal classification.

A further obstacle in this domain is the scarcity of large, publicly available, and consistently labelled datasets for construction-equipment operations. Most reported studies rely on private datasets collected from a single machine, site, or operator, which complicates benchmarking and limits the comparability and reproducibility of results across the literature [1,7,15,50,61]. Differences in sensor placement, sampling rate, window length, and activity taxonomy further hinder direct comparison between methods, and few studies report performance under a balanced, stratified evaluation protocol [20,21,41,42]. As a consequence, relatively little work has systematically examined how time-series-imaging encodings perform on multi-channel excavator operational data, and the relative contribution of the encoding step as distinct from the classifier remains underexplored.

Vision-based methods offer rich spatial detail but are fragile under real earthwork conditions, whereas conventional sensor-based methods are robust yet limited in their ability to model complex temporal dependencies, and time-series imaging has rarely been applied to excavator monitoring. The present study addresses this gap by encoding synchronised multi-sensor excavator signals as GADF images and classifying them with a convolutional network, thereby combining the visibility-independence of onboard sensing with the discriminative feature-learning capacity of image-based deep models.

3. Methodology

This section describes the proposed excavator activity recognition framework in detail. The methodology comprises five principal stages: (i) acquisition of synchronized multi-sensor operational data from the excavator; (ii) definition and labelling of the target activities; (iii) preprocessing and normalization of the raw sensor streams; (iv) transformation of the segmented time-series signals into GADF images; and (v) classification of the resulting representations using a convolutional neural network. The overall workflow is illustrated in Figure 1, and each stage is elaborated in the following subsections.

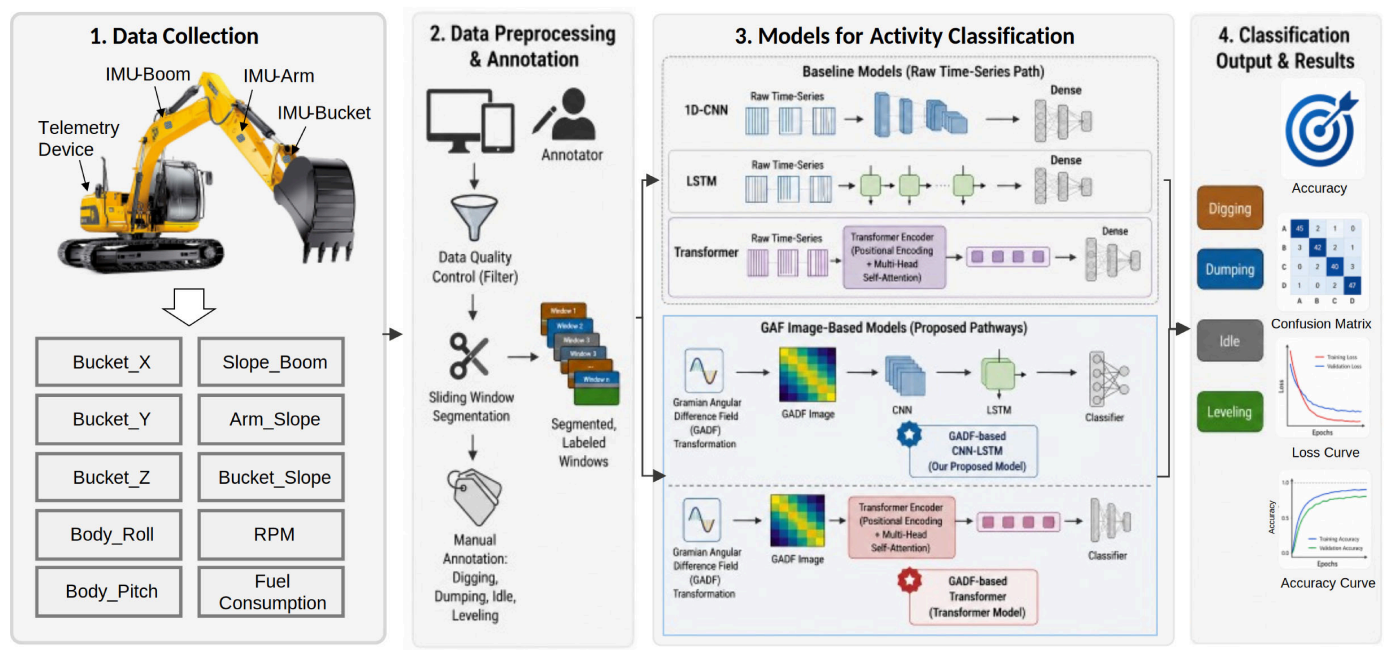


Figure 1. Overall workflow of the proposed GAF-based excavator activity recognition framework, from multi-sensor data acquisition to CNN-LSTM classification.

3.1. Sensor Data Collection

Operational data were collected from a DEVELON DX225LC-7 hydraulic excavator equipped with IMU sensors and an onboard machine control and telematics monitoring system during earthwork operations. Data were recorded at a sampling frequency of 50 Hz over a two-day collection period involving a single operator under consistent daylight conditions and broadly similar earthwork operating scenarios. The data acquisition system continuously recorded machine kinematics, attachment motion, body orientation, and engine operating conditions throughout the excavation process. Unlike vision-based approaches, which rely on external cameras and are susceptible to environmental conditions, the proposed framework utilizes onboard sensor measurements directly obtained from the machine, ensuring reliable data collection under practical construction site environments.

The collected dataset consists of synchronized multi-sensor time-series measurements representing the dynamic behavior of the excavator during operation. The primary sensor variables include bucket positional coordinates in the machine coordinate system (bucket_x, bucket_y, and bucket_z), boom angle (slope_boom), arm angle (slope_arm), bucket angle (slope_bucket), machine body roll (body_roll), machine body pitch (body_pitch), engine revolutions per minute (RPM), and fuel consumption rate. These variables collectively characterize both the geometric motion of the excavator attachment and the operational state of the machine. The instrumented excavator and representative recordings of these channels are shown in Figure 2.

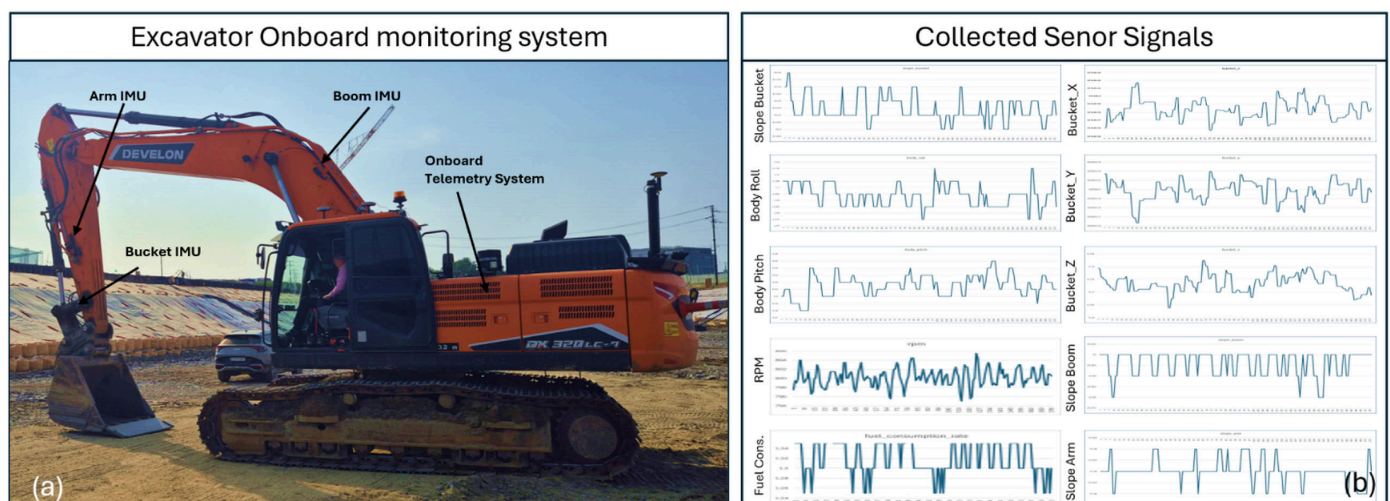


Figure 2. Excavator instrumentation and collected sensor data. (a) Excavator equipped with an onboard telemetry system and IMUs for data acquisition. (b) Representative multivariate time-series signals recorded during excavator operation. Sensor channels include bucket position (m), boom, arm, and bucket angles ($^{\circ}$), body roll and pitch ($^{\circ}$), engine RPM (rpm), and fuel rate (L/h).

Bucket positional coordinates provide direct information regarding the spatial trajectory of the excavation tool, while the boom, arm, and bucket angles describe the kinematic configuration of the excavator linkage system. Body roll and pitch measurements capture machine orientation changes that may occur during excavation and levelling activities. Engine RPM and fuel consumption rate provide indirect indicators of machine workload and operational intensity. Together, these measurements enable comprehensive characterization of excavator activities.

All sensor signals were automatically synchronized through the machine control system using a common timestamp reference. Data were recorded at a sampling frequency of 50 Hz. Data collection was conducted over two days under a single operator and consistent daylight conditions, with broadly similar operating scenarios maintained across

sessions. From this continuous record, representative segments corresponding to each activity class were extracted to construct the labelled dataset used for GADF encoding. Since the excavator remained stationary during data collection, global positioning variables such as latitude, longitude, projected coordinates, and heading angle were excluded from further analysis because they contributed minimal information for activity discrimination. Instead, the study focuses on attachment motion and machine-state variables that directly reflect operational behavior. The resulting labelled multi-sensor dataset forms the basis for the subsequent preprocessing, GADF transformation, and deep learning-based activity recognition framework.

3.2. Activity Definition

The collected dataset encompasses four representative excavator activities commonly observed in earthwork operations: digging, dumping, levelling, and idle states. These activities constitute the primary components of a typical excavation cycle and exhibit distinct motion patterns in the excavator attachment, machine posture, and engine operating conditions.

Digging refers to the excavation process in which the bucket penetrates the ground surface, collects soil, and lifts the material from the excavation area. During this activity, significant variations are observed in bucket position, bucket angle, arm angle, and boom angle due to coordinated movements of the excavator linkage system. Engine RPM and fuel consumption typically increase as the machine experiences higher operational loads.

Dumping occurs when the excavator releases the excavated material at a designated dumping location. This activity is characterized by upward boom movement, outward arm extension, and rapid bucket rotation to discharge the material. Compared with digging, dumping exhibits distinctive bucket-angle trajectories and different attachment kinematic patterns.

Levelling involves grading or smoothing the ground surface using controlled bucket movements. Unlike digging and dumping, levelling is generally characterized by repetitive and relatively smooth attachment motions with smaller vertical displacements. The bucket remains close to the ground surface while maintaining continuous contact for surface finishing operations.

Idle represents periods during which the excavator is operational but not actively performing earthmoving tasks. During idle states, bucket position and attachment angles remain relatively stable with minimal movement. Engine RPM and fuel consumption remain at lower levels compared with active working states.

The collected sensor data were segmented and manually labelled according to these four activity categories by reviewing synchronized video recordings of excavator operation and assigning each segment to the corresponding predefined activity. Each time segment was assigned a single activity label corresponding to the dominant operational behavior observed during that interval. The resulting labelled dataset was subsequently utilized for supervised training and evaluation of the proposed GAF-CNN-LSTM activity recognition framework.

Activity labels were assigned by multiple annotators through timestamp-synchronized review of the corresponding video recordings, strictly following the activity definitions specified in Section 3.2. To verify labeling consistency, a subset of labeled sections was independently cross-checked between annotators. Transition boundaries between activities were defined according to activity-specific kinematic criteria rather than annotator judgment alone; for example, the dumping activity was restricted to the specific interval in which the bucket was actively releasing material through an outward rotation motion, with adjacent repositioning or approach movements excluded from this label.

3.3. Data Preprocessing

Prior to GADF transformation and deep learning model training, a structured preprocessing pipeline was implemented to improve data quality, reduce noise, and ensure consistency among the collected sensor signals. The preprocessing procedure consisted of activity labeling, data quality control, signal normalization, and temporal segmentation.

The collected excavator operational data were first reviewed and segmented according to the predefined activity classes described in Section 3.2, namely digging, dumping, levelling, and idle. Activity labels were assigned through review of the synchronized video recordings, as described in Section 3.2, according to the dominant machine behavior within each time interval. These labels served as ground truth for supervised learning.

To improve data reliability, quality control procedures were applied to the multivariate sensor dataset. Sensor records containing missing values, duplicated timestamps, corrupted measurements, or abnormal outliers resulting from communication errors were removed. Since multiple sensors operated simultaneously, timestamp verification was performed to ensure temporal consistency across all recorded variables. Following data cleaning, the remaining sensor streams were synchronized using their common timestamp reference to generate a unified multivariate time-series dataset.

Because the collected sensor variables possess different physical units and numerical ranges, including bucket position coordinates, joint angles, body orientation, engine RPM, and fuel consumption rate, normalization was required prior to GADF encoding. Rather than normalizing across the full recording, normalization was applied locally at the sub-window level as part of the GADF transformation procedure, described in Section 3.4.

The continuous multivariate sensor data corresponding to each labelled activity were divided into fixed-length, non-overlapping windows of 150 samples (3 s at 50 Hz), with each window drawn entirely from a single activity sequence to ensure no window spans a transition between activities. This procedure enables the extraction of local operational patterns while increasing the number of training samples available for model learning.

The resulting segmented and normalized multivariate time-series data were subsequently utilized as inputs for the GADF transformation described in the following section. The multivariate sensor sequences within each time window were transformed into a single image-based representation that preserves temporal relationships and inter-sensor correlation structures, enabling effective feature extraction through convolutional neural networks.

3.4. GADF Transformation

The central component of the proposed framework is the transformation of one-dimensional multivariate sensor signals into two-dimensional image representations using the GAF encoding. Of the two GAF variants, the summation field (GASF) and the difference field (GADF), this study uses the GADF. Throughout this paper, “GAF” denotes the general encoding family and is retained in the model names GAF-CNN-LSTM and Transformer-GAF, whereas “GADF” denotes the specific matrices and images generated in this study. This transformation allows operational time-series data, which are conventionally processed using sequential models, to be analysed using image-based convolutional neural networks that are highly effective at extracting local and hierarchical spatial features.

Each 150-sample activity window was further divided into 15 non-overlapping sub-windows of 10 time steps. Within each sub-window, sensor values were first Min–Max normalized per channel to the [0, 1] range:

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

where $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values of that sensor channel within the sub-window. The resulting normalized sub-window was transposed and flattened into a univariate sequence of length $L = 10 \times 10 = 100$ by concatenating the ten sensor channels.

When a sensor channel is constant over a window ($x_{\min} = x_{\max}$), the denominator equals zero; in this case, the normalized value is set to 0.5 to avoid division by zero, which maps to $\varphi = \pi/2$ after rescaling to $[-1, 1]$; the channel's own diagonal block of the GADF is then uniformly zero, and its cross-channel entries reduce to $\pm \cos \varphi_j$ of the other channels, so the constant channel contributes no spurious texture of its own. Given a normalized sensor sequence within a temporal window, $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n\}$, each value is first represented in a polar coordinate system. The sequence is internally rescaled to $[-1, 1]$ by the GADF transformation prior to angular encoding, as required for the mapping to be well-defined. The value can be encoded as an angular cosine, while the corresponding time index is encoded as the radius. The angle associated with each sample is obtained as:

$$\phi_i = \arccos(\tilde{x}_i), \quad -1 \leq \tilde{x}_i \leq 1 \quad (2)$$

This angular encoding establishes a bijective mapping between the normalized signal amplitude and an angle in the range $[0, \pi]$, preserving the relative amplitude ordering; however, the subsequent RGB colormap quantization introduces a degree of irreversibility, and the full pipeline should be regarded as near-lossless rather than strictly invertible. After the angular representation is obtained, the Gramian Angular Difference Field (GADF) is constructed by computing the sine of the angular difference between every pair of time points i and j :

$$G_{i,j} = \sin(\phi_i - \phi_j) \quad (3)$$

This sequence was transformed via the GADF formulation into a single 100×100 matrix per sub-window, in which diagonal blocks capture each sensor's temporal self-correlation and off-diagonal blocks capture cross-sensor angular relationships, following the multi-sensor concatenation strategy of Zhao et al. [62]. The resulting single-channel matrix was rendered as a three-channel color image using a fixed colormap for compatibility with the convolutional network's input format; the three channels therefore represent a colormap encoding of one underlying value per pixel rather than three independently informative signals. Each 150-sample window thus yields a sequence of 15 RGB images, which together form the input sequence to the CNN-LSTM architecture described in Section 3.5.

The preservation of temporal relationships is a key advantage of the GADF encoding. Since the matrix is constructed from pairwise angular differences between temporally ordered observations, each element reflects the relative temporal relationship between two time instants while implicitly preserving correlations among the multiple sensor variables. Cyclic, repetitive, and transient operational patterns such as coordinated boom, arm, and bucket movements during digging or the smooth, low-amplitude motions observed during levelling are therefore represented as distinctive spatial textures within the GADF image. These texture patterns provide highly discriminative visual features that facilitate activity recognition.

Each generated GADF image serves as a compact visual representation of the multivariate sensor data within one sub-window and is used directly as the input to the convolutional neural network described in Section 3.5. By jointly encoding the temporal dynamics of all ten sensors into a single image, the proposed representation enables the network to learn both intra-sensor temporal patterns and cross-sensor relationships without requiring separate image generation or channel stacking. The complete encoding

process, from the one-dimensional sensor signal to the resulting two-dimensional image, is illustrated in Figure 3.

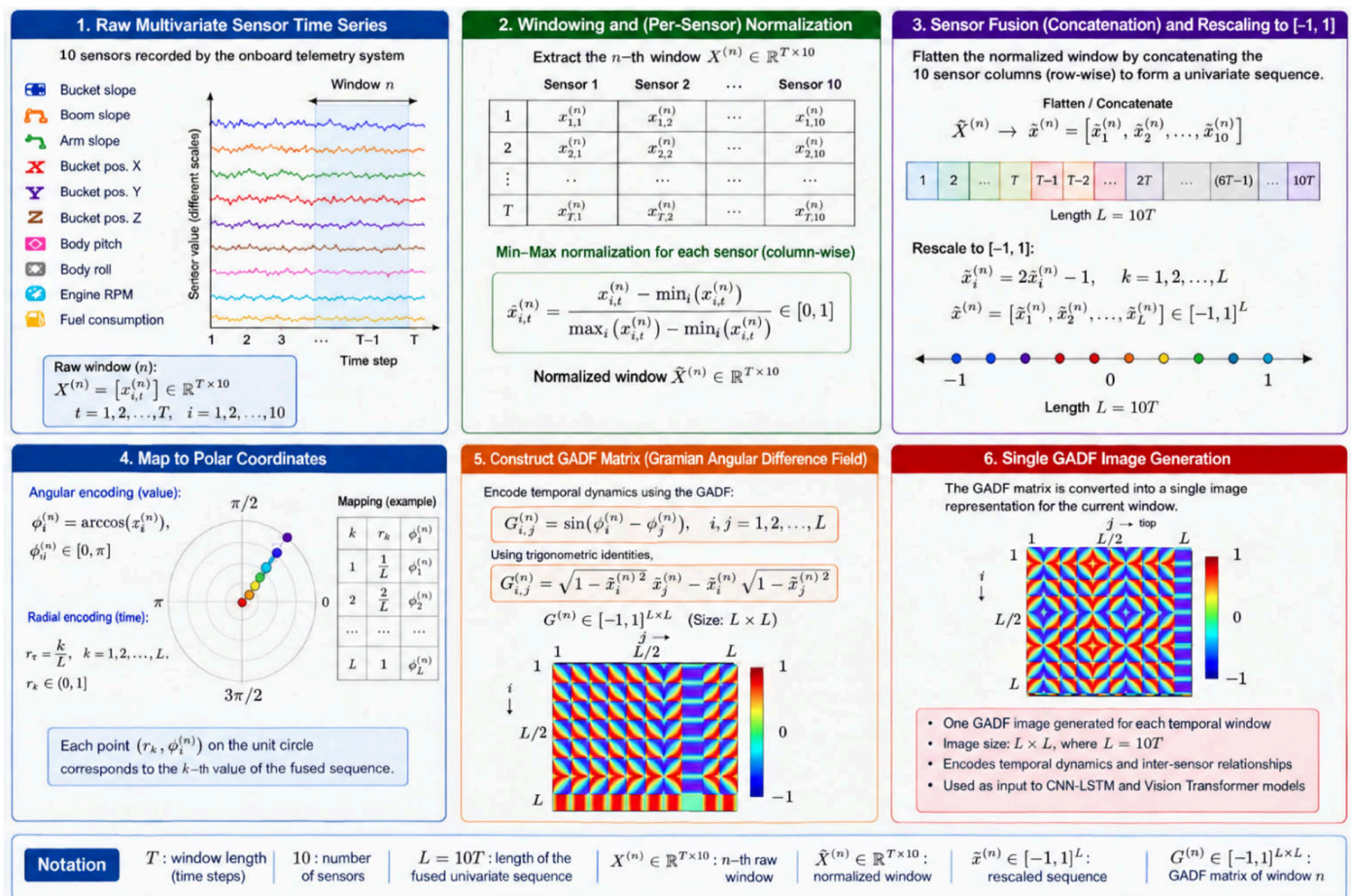


Figure 3. Illustration of the Gramian Angular Difference Field (GADF) encoding of the ten excavator sensor channels within one window into a single two-dimensional image.

The GADF representation is naturally compatible with convolutional architectures. Convolutional filters extract local texture-like features from the encoded temporal correlation patterns; pooling operations may provide some tolerance to small temporal shifts and measurement noise, and successive convolutional layers learn increasingly abstract representations of excavator operating behaviors. Consequently, the proposed GADF-based CNN framework is intended to capture discriminative motion characteristics that are difficult to extract directly from raw multivariate time-series data.

3.5. CNN-LSTM Architecture

Once the multi-sensor signals are encoded into GADF image sequences, a Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture is employed to learn both spatial and temporal representations of the excavator operations. The convolutional neural network extracts discriminative spatial features from each GADF image, while the LSTM network captures the temporal dependencies among consecutive operational windows. This hybrid architecture enables the model to jointly exploit the spatial characteristics encoded in the GADF images and the temporal evolution of excavator activities.

The network receives an input tensor of size $T \times C \times N \times N$, where T denotes the sequence length, C represents the number of image channels, and $N \times N$ corresponds to the spatial resolution of each GADF image. In this study, each input sample consists of a se-

quence of 15 RGB GADF images ($T = 15, C = 3$), with each image resized to 100×100 pixels prior to feature extraction.

The convolutional feature extractor consists of two convolutional blocks. The first block comprises a 3×3 convolutional layer (32 filters, stride 1, padding 1), followed by a ReLU activation function, 2×2 max-pooling (stride 2) for spatial down-sampling, and a two-dimensional dropout layer for regularization. The second block comprises a 3×3 convolutional layer (64 filters, stride 1, padding 1), followed by a ReLU activation function, adaptive average pooling to a fixed 4×4 spatial grid, and a two-dimensional dropout layer. The resulting $64 \times 4 \times 4$ feature map is flattened and passed through a linear projection layer to produce a compact 128-dimensional kinematic feature vector F_{GAF} for each image in the sequence. The successive convolutional blocks progressively learn hierarchical feature representations, ranging from low-level texture and edge information to more discriminative operational patterns associated with different excavator activities. Unlike conventional image classification networks, the extracted feature maps are projected into a fixed-dimensional representation and preserved for temporal modelling rather than being directly classified, as shown in Figure 4.

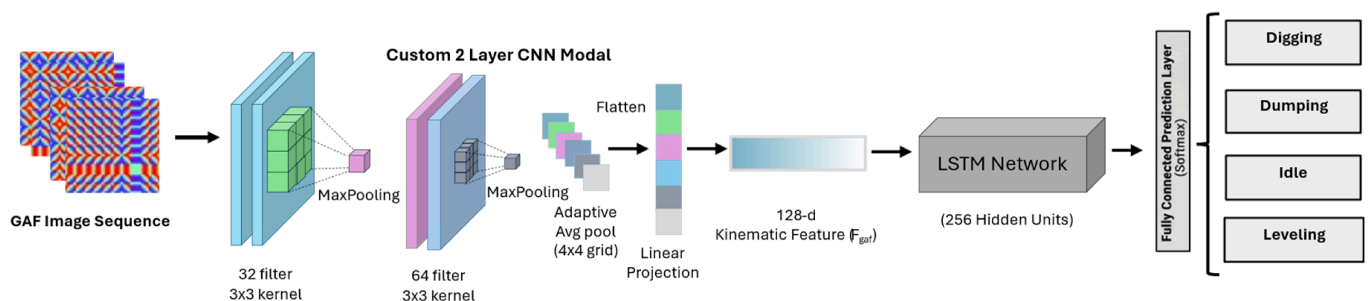


Figure 4. Overall architecture of the proposed GAF-CNN-LSTM framework for excavator activity recognition using GADF image sequences.

The sequence of feature vectors generated by the CNN is subsequently provided to an LSTM network comprising a single layer with 256 hidden units, which models the temporal relationships between consecutive GADF images. The LSTM processes the sequence sequentially and retains relevant contextual information through its memory cells, enabling the network to capture temporal variations occurring during different excavation operations. The hidden representation corresponding to the final time step is selected as the overall descriptor of the input sequence.

The final temporal feature vector is passed through a fully connected classification layer preceded by a dropout layer to reduce overfitting during training. The classifier produces raw prediction scores for each activity class, while the softmax operation is implicitly incorporated within the cross-entropy loss function to obtain the probability distribution over the target excavator activities. The activity associated with the highest posterior probability is selected as the predicted class.

3.6. Training and Hyperparameters

The proposed CNN-LSTM model was trained in a supervised manner using the labelled GADF image sequences. During preprocessing, all input images were resized to 100×100 pixels and normalized to the range $[0, 1]$. To improve the robustness of the model and reduce overfitting, Gaussian noise augmentation was applied to the training samples, while the validation dataset was evaluated without augmentation. PCA-based feature visualization, presented in Section 5, was applied solely to the LSTM embeddings for interpretability purposes; PCA was fitted exclusively on the training-set features, and

the resulting projection was applied to validation-set features without refitting, ensuring no information leakage.

The network was optimized using the Adam optimizer together with the multi-class cross-entropy loss function. A Reduce-on-Plateau learning-rate scheduler was employed to automatically decrease the learning rate when the validation loss ceased to improve, thereby facilitating stable convergence during training. The model was trained for a maximum of 40 epochs using mini-batches, and the latest network parameters were saved after each training epoch. Among these checkpoints, the model corresponding to the epoch with the minimum validation loss was selected as the final model used for all reported evaluation results, rather than the final-epoch or an early-stopped model. The principal training configuration and hyperparameters are summarized in Table 1.

Table 1. Training configuration and hyperparameters of the proposed GAF-CNN-LSTM model.

Parameter	Value
Input image size	100 × 100 pixels
Number of input channels	3 (RGB GADF)
Sequence length	15 images
Convolution kernel size	3 × 3
Number of CNN blocks	2
CNN filters	32, 64
Activation function	ReLU
Pooling operation	Max Pooling (block 1) + Adaptive Avg Pooling 4 × 4 (block 2)
Linear projection	128-d
CNN Dropout	0.30
LSTM layers	1
LSTM hidden units	256
Fully Connected Dropout	0.40
Output classes	4 (Digging, Dumping, Idle, Levelling)
Loss function	Cross-Entropy Loss
Optimizer	Adam
Initial learning rate	0.0005
Learning rate scheduler	ReduceLROnPlateau (factor = 0.5, patience = 10)
Batch size	16
Number of epochs	40
Weight decay	1×10^{-5}
Convolution kernel size	3 × 3
Stride	1
Padding	1
Pooling operation	Max Pooling (2 × 2)
CNN Dropout	0.30
Fully Connected Dropout	0.40
Weight decay	1×10^{-5}

4. Experimental Setup

The proposed framework was implemented in Python (version 3.14.6) using the PyTorch deep learning library (version 2.14.0). All experiments were conducted on an HP

Z8 G4 workstation running Windows 10 (64-bit), equipped with dual Intel Xeon Gold 6242R CPUs and an NVIDIA RTX GPU, with CUDA acceleration used for both model training and inference. The same hardware and software environment was employed throughout all experiments to ensure a fair and consistent evaluation.

The labelled dataset was constructed from the segmented and GADF-encoded operational windows described in Section 3. The samples were distributed across the four activity classes as summarized in Table 2. To prevent data leakage between training and validation, the split was performed at the operation-sequence level rather than at the individual window level: each labelled activity sequence was assigned entirely to either the training set or the validation set, ensuring that no sequence—and therefore no window derived from it—appeared in both subsets. Class proportions were preserved across both subsets, giving 952 training and 168 validation sequences (238 and 42 per class, respectively).

Table 2. Distribution of GADF image sequences across the four excavator activity classes.

Activity	Number of Samples	Proportion
Digging	280	25.0%
Dumping	280	25.0%
Levelling	280	25.0%
Idle	280	25.0%
Total	1120	100%

It should be noted that the 1120 samples represent complete GADF image sequences—each comprising 15 images derived from one non-overlapping 150-sample window—and are not individual frames. The equal class distribution of 280 sequences per class reflects the balanced nature of the data collection protocol, in which recording sessions were conducted to ensure sufficient coverage of all four activity classes without artificial oversampling or undersampling.

Model performance was assessed using four standard classification metrics: accuracy, precision, recall, and the F1-score. Accuracy measures the overall proportion of correctly classified windows, while precision and recall characterize the per-class reliability and completeness of the predictions. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced measure that is robust to mild class imbalance. Precision, recall, and the F1-score were computed per class and then macro-averaged across all activities. The metrics are defined as follows:

$$\text{Accuracy} = (\sum_c \text{TP}_c) / N$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

where TP_c is the number of correctly classified windows of class c , N is the total number of evaluated windows, and TP , FP , and FN in the precision and recall expressions are computed per class in a one-versus-rest manner.

5. Results and Discussion

This section presents the experimental results of the proposed Gramian Angular Field CNN-LSTM (GAF-CNN-LSTM) framework for excavator activity recognition. The evaluation covers classification performance across four activity classes, comparison with competing baseline methods, and an interpretive discussion of the observed performance gains.

5.1. Classification Performance

The proposed GAF-CNN-LSTM framework was trained over 40 epochs using a learning rate of 0.0005 and evaluated on a held-out validation set comprising 168 samples equally distributed across the four excavator activity classes: digging, dumping, idle, and levelling (42 samples each). Training progressed steadily from an initial validation accuracy of 25.00% at Epoch 1 to a peak validation accuracy of 95.24% at Epoch 36, accompanied by a corresponding training loss reduction from 1.44 to 0.055. The training curve, as shown in Figure 5, demonstrates stable learning without significant oscillation after Epoch 25, indicating that the GADF image representation provides consistent and discriminative input features for the CNN-LSTM architecture. To assess reproducibility, three independent training runs were conducted with different random seeds, yielding validation accuracies of 95.24%, 95.24%, and 92.26% with a mean of $94.25\% \pm 1.72$ percentage points and a mean macro F1-score of 0.943 ± 0.017 , indicating broadly stable performance across initializations.

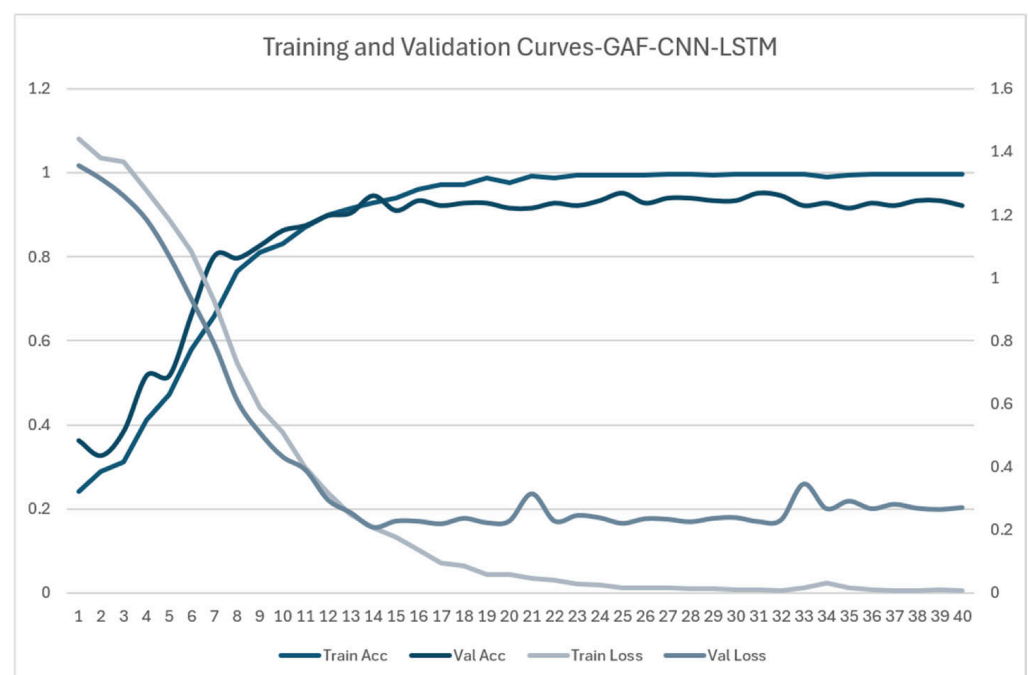


Figure 5. Training and validation accuracy and loss curves of the proposed GAF-CNN-LSTM model over 40 epochs. Accuracy values correspond to the left y-axis; loss values correspond to the right y-axis. The best model checkpoint was selected at Epoch 36 based on minimum validation loss (Val Loss = 0.1931, Val Acc = 95.24%).

To further investigate the discriminative capability of the learned feature representations, Principal Component Analysis (PCA) was applied to the feature vectors extracted from the final LSTM layer. PCA was fitted on the training-set embeddings and applied, without refitting, to the 168 validation samples (42 per class), which are the only samples plotted in Figure 6. Because the first three principal components retain only part of the variance of the 256-dimensional embeddings, the plot should be read as a qualitative illustration; class separability is quantified by the confusion matrix in Figure 7. As illustrated in Figure 6, the feature embeddings form four distinct clusters corresponding to the four excavator activities. The digging, idle, and levelling classes exhibit compact intra-class distributions with clear separation from neighbouring classes, indicating that the proposed CNN-LSTM model successfully learns discriminative feature representations. A small degree of overlap is observed between the dumping and levelling clusters, suggesting that these activities share similar temporal and kinematic characteristics during certain

operational phases. Nevertheless, the overall cluster separation indicates that the combined GADF encoding and CNN-LSTM architecture effectively transforms the multi-sensor time-series data into a feature space that facilitates accurate activity classification.

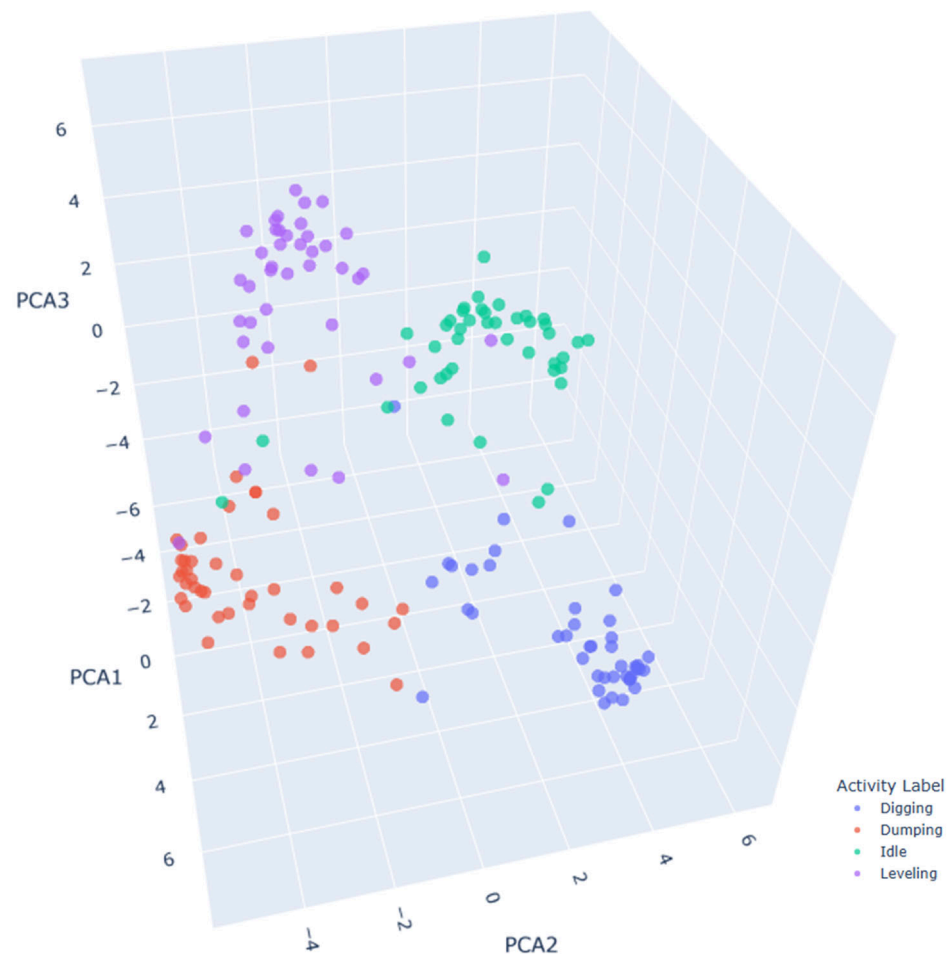


Figure 6. Three-dimensional Principal Component Analysis (PCA) visualization of the feature representations learned by the proposed CNN-LSTM model for excavator activity recognition. The plotted points are the 168 validation samples (42 per class) only, projected with a PCA fitted on the training-set embeddings. The first three components retain only part of the embedding variance, so the plot is a qualitative illustration.

The per-class classification results at peak performance are reported in Table 3. The digging class achieved the highest recall of 1.000, indicating that all digging activity windows were correctly identified without any missed detections. The idle class attained a precision of 1.000 and an F1-score of 0.976, reflecting its highly distinctive operational signature characterized by minimal attachment movement and low engine loading. The levelling class reached an F1-score of 0.941, supported by a recall of 0.952. The dumping class, which exhibited the greatest inter-class confusion, particularly with the levelling category, achieved a precision of 0.927 and an F1-score of 0.916. The overall macro-averaged F1-score across all classes was 0.952, with an overall classification accuracy of 95.24%.

The confusion matrix, as shown in Figure 7 at the best-performing epoch, reveals that the digging class achieved perfect recall (42/42 correct classifications), while two dumping samples were misclassified as digging, two dumping samples were misclassified as levelling, and two levelling samples were misclassified as dumping. These residual confusions are consistent with the kinematic similarity between dumping and levelling activities, as both involve horizontal attachment displacements without the high-torque

bucket penetration characteristic of digging. The idle class incurred two misclassifications, one as dumping and one as levelling, attributable to brief attachment repositioning motions that occur during transition periods between idle and active states.

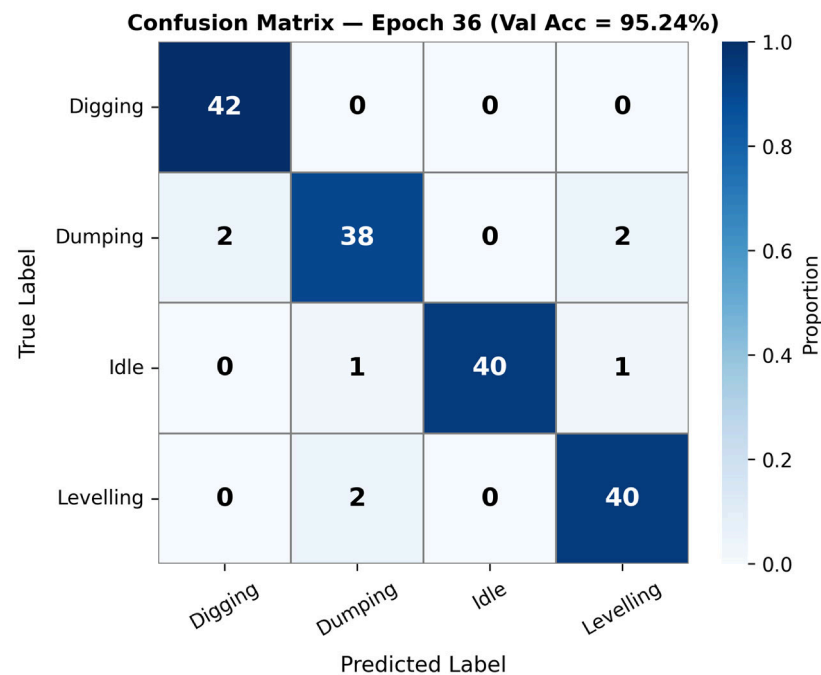


Figure 7. Confusion matrix at the best checkpoint (Epoch 36). The model achieves perfect classification for digging (42/42), with minor confusions in dumping (4 errors), idle (2 errors), and levelling (2 errors).

Table 3. Per-class classification performance of the proposed GAF-CNN-LSTM framework.

Activity	Precision	Recall	F1-Score	Support	Accuracy
Digging	0.955	1.000	0.977	42	95.24%
Dumping	0.927	0.905	0.916	42	
Idle	1.000	0.952	0.976	42	
Levelling	0.930	0.952	0.941	42	
Macro Avg	0.953	0.952	0.952	168	

5.2. Comparison with Existing Methods

To contextualize the performance of the proposed framework, four baseline approaches were evaluated on the same dataset and training/validation split: LSTM operating on raw time-series sequences, a one-dimensional convolutional neural network (1D CNN) processing raw sensor signals, a Transformer encoder applied to raw time-series data (Transformer-TS), and a Transformer encoder applied to GADF image representations (Transformer-GAF). The comparative results are summarized in Table 4.

To ensure a fair and transparent comparison, the architecture and training configuration of each baseline model are reported here. The LSTM baseline processes raw 10-channel sensor sequences of length 15, standardized using a StandardScaler fitted on the training set, through two stacked LSTM layers (hidden sizes 32 and 16), each followed by Dropout(0.3), a Dense(16, ReLU) layer, and a softmax classifier; training used Adam (lr = 0.001), batch size 16, up to 50 epochs with early stopping (patience = 10). The 1D CNN baseline uses the same raw input and comprises two convolutional blocks (Conv1D with 64 and 128 filters, kernel size 3, same padding, BatchNormalization, MaxPooling1D(2), Dropout(0.3)), fol-

lowed by GlobalAveragePooling1D, Dense layers (64 and 32 units, ReLU), and softmax output; training settings were identical to the LSTM. The Transformer-TS baseline uses the same raw input with a Dense(64) input embedding, two Transformer encoder blocks (MultiHeadAttention: 4 heads, key_dim = 64; FFN: 128 → 64; Dropout(0.3)), GlobalAveragePooling1D, Dense(64, ReLU), and softmax output; training used Adam (lr = 0.001), batch size 16, up to 50 epochs. The Transformer-GAF baseline receives the same GADF image sequences as the proposed model; each image is flattened and projected to 256 dimensions with learned positional encodings, processed by four Transformer encoder layers (8 heads, d_model = 256), mean-pooled, and classified by a linear layer; training used Adam (lr = 0.0001), batch size 8, up to 60 epochs. All baselines were evaluated on the same training/validation split with no augmentation applied to baseline inputs.

Table 4. Classification accuracy and macro F1-score comparison across all evaluated methods.

Method	Input Representation	Accuracy (%)	Macro F1	Notes
LSTM	Raw Time-Series	87.0	0.86	Baseline
1D CNN	Raw Time-Series	89.0	0.88	Baseline
Transformer-TS	Raw Time-Series	89.0	0.88	Baseline
CNN-LSTM	Raw Time-Series	92.5	0.93	Ablation
Transformer-GAF	GADF Images	94.6	0.947	Baseline
GAF-CNN-LSTM (Proposed)	GADF Images	95.24	0.952	Best

Note: CNN-LSTM (raw-input) values are means over repeated training runs; the corresponding mean for the proposed GAF-CNN-LSTM over three runs is 94.25% accuracy (macro F1 = 0.943).

Among the raw time-series baselines, the LSTM model achieved the lowest validation accuracy of 87.0% with a macro F1-score of 0.86. The 1D CNN and Transformer-TS models both reached 89.0% accuracy and a macro F1-score of 0.88, demonstrating that architectural improvements alone without changing the input representation yield only modest gains. When a Transformer encoder was applied to GADF image inputs (Transformer-GAF), performance improved substantially to 94.6% accuracy and a best-epoch macro F1-score of 0.947, suggesting that GADF encoding contributes to the observed performance improvement, although the Transformer-TS and Transformer-GAF models differ in architecture (number of encoder layers, attention heads, and embedding dimension) and in training settings (learning rate, batch size, and epoch budget) as well as in input representation, so this gain cannot be attributed to GADF encoding alone.

To examine the contribution of the architecture separately from that of the input representation, the CNN-LSTM architecture was also evaluated on raw time-series input across repeated training runs, achieving a mean accuracy of 92.5% and a mean macro F1-score of 0.93 across runs. This substantially outperforms the raw-input LSTM, 1D CNN, and Transformer baselines, suggesting that the CNN-LSTM architecture itself accounts for much of the performance gain over conventional raw-input methods. Averaged over repeated training runs, accuracy increases from 92.5% for the raw-input CNN-LSTM to 94.25% for the GAF-CNN-LSTM. This comparison is indicative rather than controlled. The raw-input model requires a different convolutional input stage to process one-dimensional sequences, and its training settings, augmentation, and model-selection procedure were not controlled to match those of the GADF-input model. In addition, the 1.75-percentage-point difference is of the same order as the run-to-run standard deviation of the proposed model (1.72 percentage points). The comparison therefore does not establish the size of the contribution of GADF encoding.

The proposed GAF-CNN-LSTM achieved the highest overall accuracy of 95.24% and a macro F1-score of 0.952, surpassing the best raw time-series baseline (89.0%) by approximately 6.2 percentage points. The comparison between Transformer-GAF (94.6%) and GAF-CNN-LSTM (95.24%) is consistent with the view that the CNN-LSTM architecture, which combines local spatial feature extraction through convolutional layers with sequential context modeling through LSTM units, provides a complementary advantage over pure attention-based processing of GADF images. It should be noted, however, that the 0.64 percentage-point margin between GAF-CNN-LSTM (95.24%) and Transformer-GAF (94.6%) corresponds to approximately one correct prediction among 168 validation samples and falls within the observed run-to-run variability of the model; a larger independent test set would be required to establish the statistical significance of this difference.

5.3. Why GADF Encoding May Improve Classification Performance

The performance gap between the GADF-input models and the raw time-series baselines—approximately 5.6–6.2 percentage points relative to the best raw In Proceedings of the baseline—motivates a discussion of the mechanisms through which GADF encoding may improve excavator activity recognition. The mechanisms below are offered as explanations consistent with the results, not as experimentally verified effects.

Preservation of temporal correlation structure. Unlike direct sequential processing, GADF transformation maps a normalized time-series signal into a skew-symmetric image matrix where element (i, j) encodes the angular difference in the signal values at time steps i and j , given by $GADFi,j = \sin(\varphi_i - \varphi_j)$. This encoding explicitly captures pairwise temporal interactions across all time lags simultaneously, whereas LSTM and Transformer models must learn these dependencies implicitly through sequential hidden state updates. For excavator activities, which exhibit characteristic attachment trajectories spanning multiple time steps within a work cycle, the explicit encoding of temporal co-occurrence patterns within the GADF image provides rich structural information that is more directly accessible to convolutional feature extractors.

Spatial feature extraction from the unified multi-sensor GADF image. The ten excavator sensor channels—bucket coordinates, boom, arm, and bucket angles, body roll and pitch, engine RPM, and fuel consumption rate—are jointly encoded into a single unified GADF image, rather than being represented as separate per-sensor images or stacked into a multi-channel tensor. This representation allows the CNN to learn spatial correlations across the regions of the unified image that correspond to different sensor pairs, which may capture multi-sensor co-activation patterns characteristic of each activity class, though this study does not isolate individual sensor contributions to confirm this mechanism directly. For example, the simultaneous angular displacement of the boom and arm sensors during digging produces a distinctive spatial signature within the unified GADF image that can be detected through convolutional filtering, whereas the same pattern is more difficult to extract from interleaved raw scalar sequences.

Possible noise tolerance through image-domain aggregation (not tested experimentally). Individual sensor measurements are subject to quantization noise, transient disturbances from terrain irregularities, and operator variability. In the GADF image domain, a perturbation at time step i propagates to all elements in row i and column i of the GADF matrix (up to $2N - 1$ elements), owing to the pairwise angular difference structure. If the perturbed value becomes the new minimum or maximum of its channel within a sub-window, the Min-Max step also changes that channel's other normalized values, extending the effect to that channel's rows and columns but not beyond them. While this is broader than point-local noise, the corruption remains structurally bounded and does not redistribute across the full image, which could limit the effect of an isolated disturbance;

whether it does so was not tested in this study. This remains a theoretical characterization; formal noise-robustness analysis is left as future work. Convolutional operations with spatial pooling may therefore offer some tolerance to such structured row-and-column disturbances, as the aggregation over spatial neighborhoods could suppress noise while preserving the global pattern structure corresponding to the activity class. This remains a theoretical mechanism rather than an empirically validated property in this study; the lower misclassification rates observed for digging and idle are consistent with this hypothesis but do not isolate noise tolerance from other contributing factors. Direct evaluation through controlled noise injection or sensor-channel ablation is identified as an important direction for future work.

Enhanced inter-class separability. The confusion between dumping and levelling, which was the dominant source of error across all evaluated methods, was reduced from 5 to 10 misclassifications in the time-series baselines to 4 in the proposed method (two dumping windows predicted as levelling and two levelling windows predicted as dumping; Figure 7). This reduction suggests that the GADF representation provides improved inter-class separability for activity pairs that share similar instantaneous kinematic ranges but differ in their temporal trajectory patterns. The LSTM component of the CNN-LSTM architecture further reinforces this by modeling the sequential ordering of spatial features extracted from successive GADF image windows, enabling the model to distinguish activities based on their temporal progression rather than static snapshot features alone.

5.4. Practical Implications

The results of this study suggest that the proposed GAF-CNN-LSTM framework is a promising candidate for excavator activity monitoring in construction environments, pending validation under broader operational conditions. Several practical implications follow from these findings.

For intelligent and autonomous construction equipment, activity classification could in future inform onboard control, for example, by adapting hydraulic pressure limits or engine torque curves to the detected activity state; such closed-loop use would require validated real-time performance on embedded hardware, which was not evaluated in this study.

For fleet management and productivity monitoring systems, the proposed framework provides activity-level telemetry that can be streamed from onboard telematics units to remote management platforms with low computational overhead, since GADF encoding and CNN-LSTM inference are expected to be computationally lightweight, though inference latency was not formally measured in this study on embedded processors. Because the approach relies on onboard sensor signals rather than external cameras, it is not subject to the visibility constraints of camera-based systems, although its operation under the connectivity conditions of active earthwork sites has not yet been evaluated.

For digital twin integration, the activity labels generated by the framework could, once validated for real-time use, drive physics-based simulation models that track the operational state of the excavator, potentially supporting proactive scheduling of maintenance interventions and cycle-time optimization. The consistent detection of idle states achieved with 95.2% recall in this study is particularly valuable for identifying unproductive machine time and improving overall equipment effectiveness.

From a deployment perspective, the sensor signals utilized in this study—bucket position, attachment angles, body orientation, engine RPM, and fuel rate—are standard outputs of modern excavator machine control systems and are available on many current-generation hydraulic excavators equipped with telematics; however, field deployment would require further validation across machine types and operating environments. Where

these signals are already provided by factory-fitted machine control and telematics systems, the framework could be integrated with little additional hardware; in this study, however, IMUs were installed on the excavator for data acquisition, so the hardware requirements of a production deployment remain to be established.

6. Limitations

Although the proposed framework demonstrates strong performance, several limitations should be acknowledged to delineate the boundaries of the present study and to guide future research. First, the dataset was collected from a single hydraulic excavator, operated by a single operator, under a limited range of site conditions. Consequently, the learned operational signatures may be specific to the kinematics, hydraulic response, and telematics characteristics of this particular machine, and the generalizability of the model to excavators of different sizes, manufacturers, and control configurations remains to be verified. Cross-machine and cross-site validation on larger and more heterogeneous datasets is therefore an important direction for future work.

Second, the recognition task was restricted to four representative activities—digging, dumping, levelling, and idle. Real earthwork operations comprise additional behaviours, such as travelling, truck loading, ground compaction, and combined or transitional actions, that were not represented in the current dataset. Moreover, because the excavator remained stationary during data acquisition, global positioning and heading variables were excluded; the framework has therefore not yet been evaluated under machine-travel conditions. Extending the activity taxonomy and incorporating mobility scenarios would broaden the practical applicability of the approach.

Third, the method depends on the availability and reliability of onboard operational sensors. Although modern excavators are increasingly equipped with machine-control and telematics systems, the approach presumes access to synchronized, well-calibrated sensor streams. Sensor faults, signal drift, missing channels, or differences in sampling configuration across equipment fleets could degrade recognition performance, and the robustness of the GAF-CNN-LSTM framework to such degraded or partially missing inputs was not systematically examined in this study.

Fourth, environmental and operational variability—including soil type, terrain irregularity, weather, and individual operator behaviour—was only partially captured by the collected data, and its influence on recognition accuracy warrants dedicated investigation. In addition, activity labels were assigned manually through review of synchronized video recordings, which may introduce label noise at the transition boundaries between activities. Finally, the framework was validated offline on a held-out evaluation set; its real-time inference latency and computational footprint on embedded or edge hardware, as well as its behaviour in a live online-deployment setting, remain to be benchmarked. Addressing these aspects would further strengthen the practical readiness of the proposed system.

Fifth, this study evaluated the model using a training/validation split alone; no independent test set, separate from the data used for model selection, was reserved. While leakage between training and validation was avoided through sequence-level splitting, reported performance may still be optimistic relative to deployment on entirely unseen operation sequences. Furthermore, the validation set served a dual role in this study—it was used both by the ReduceLROnPlateau learning-rate scheduler during training and as the basis for selecting and reporting the best-checkpoint performance—which may introduce a degree of optimism in the reported results. Future work should incorporate a held-out test set, ideally spanning additional operators, sites, and time periods, to obtain an unbiased estimate of generalization performance. In addition, neither the comparison between Transformer-TS and Transformer-GAF nor that between the raw-input and GADF-input

CNN-LSTM was conducted under matched configurations; a fully controlled ablation, holding the architecture (apart from the input stage), training hyperparameters, augmentation, and model-selection procedure constant and varying only the input representation, is required to isolate the effect of GADF encoding, and is left to future work.

7. Conclusions

This study presented a GAF-based multi-sensor encoding framework for excavator activity recognition, applying a known image-encoding and CNN classification approach to a new set of sensors and application area. By transforming synchronized time-series signals—bucket position, boom, arm, and bucket angles, body roll and pitch, engine RPM, and fuel consumption—into two-dimensional image representations, the proposed GAF-CNN-LSTM framework recasts the activity-recognition problem as an image-classification task in which convolutional networks can exploit the temporal correlation structure preserved by the GADF encoding. Evaluated on four representative earthwork activities (digging, dumping, levelling, and idle), the framework achieved a classification accuracy of 95.24% and a macro-averaged F1-score of 0.952, outperforming conventional raw time-series baselines including LSTM, 1D CNN, and Transformer models by 6.24–8.24 percentage points.

These results indicate that GADF-based image encoding provided a more discriminative representation of excavator operational behaviour than direct sequential processing in the evaluated dataset, while relying solely on onboard sensor signals of the kind available on many telematics-equipped machines; its robustness to sensor noise was not tested and is not claimed. The proposed approach thus bridges the complementary weaknesses of vision-based methods, which are sensitive to occlusion, dust, and illumination, and of conventional sensor-based methods, which struggle to capture complex temporal dependencies. The framework offers a promising foundation for real-time activity monitoring in intelligent construction equipment, though validation to date is limited to a single excavator and operator; broader deployment in fleet-productivity analysis and digital-twin integration will require cross-machine, cross-site, and cross-operator validation, as outlined in Section 6. Future work will extend the activity taxonomy, validate the method across multiple machines and construction sites, and investigate real-time embedded deployment to further advance sensor-driven monitoring for smart construction.

Author Contributions: Conceptualization, A.S.; methodology, A.S. and A.U.; software, A.U., W.A.T. and S.A.; validation, A.U., W.A.T. and S.A.; formal analysis, A.S.; investigation, A.U., W.A.T. and S.A.; data curation, A.U.; writing—original draft preparation, A.S. and A.U.; writing—review and editing, A.S., W.A.T. and S.A.; visualization, A.U., W.A.T. and S.A.; project administration, W.A.T.; funding acquisition, W.A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, grant number KF265563.

Institutional Review Board Statement: The research was conducted with integrity, fidelity, and honesty. All ethical procedures were considered.

Data Availability Statement: The dataset used and/or analyzed during the current study is available from the corresponding author upon reasonable request.

Acknowledgments: The authors gratefully acknowledge the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, for its support of this work. The contribution of A.S. (Abubakar Sharafat) to this research was made before June 2026, while he was affiliated with Hanyang University.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

References

- Sharafat, A.; Latif, K.; Deng, T.; Seo, J. Excavator activity recognition under occlusion via multi-camera deep learning. *Results Eng.* **2026**, *29*, 108611. [\[CrossRef\]](#)
- Benrouba, F.; Sharafat, A.; Ullah, A.; Seo, J. Autonomous productivity measurement NDT-SLAM-based algorithm for excavator earthwork operations. *KSCE J. Civ. Eng.* **2026**, *30*, 100483. [\[CrossRef\]](#)
- Ma, Z.; Wang, G.; Wang, Y. Experimental Research on the Dynamic Characteristics of Several Model Soils in Small Strain Range and Application in Shaking Table Model Test. *Buildings* **2023**, *13*, 592. [\[CrossRef\]](#)
- Nie, Q.; Wu, B.; Wang, Z.; Dai, X.; Chen, L. Incorporation of Disposed Face Mask to Cement Mortar Material: An Insight into the Dynamic Mechanical Properties. *Buildings* **2024**, *14*, 1063. [\[CrossRef\]](#)
- Nie, Q.; Zhang, J.; Wang, Y.; Wang, D.; Wang, Q.; Shi, Y.; Wang, C. Physics-Informed Machine Learning for Predicting Stress Wave Transmission Across Realistic Rock Joints. *IEEE Access* **2025**, *13*, 212735–212744. [\[CrossRef\]](#)
- Tanoli, W.A.; Ullah, A.; Sharafat, A.; Ismaeil, E.M.H. A Multi-Model BIM-Based Framework for Integrated Digital Transformation of Design to Construction of Large Complex Underground Caverns. *Buildings* **2025**, *15*, 2834. [\[CrossRef\]](#)
- Kim, I.S.; Latif, K.; Kim, J.; Sharafat, A.; Lee, D.E.; Seo, J. Vision-Based Activity Classification of Excavators by Bidirectional LSTM. *Appl. Sci.* **2023**, *13*, 272. [\[CrossRef\]](#)
- Kim, J.; Chi, S. Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. *Autom. Constr.* **2019**, *104*, 255–264. [\[CrossRef\]](#)
- Zhu, Y.; Shu, J.; Ding, W.; Yue, C.; Lu, Y.; Zhang, J. A crack detection and quantification framework for high-resolution images using Mamba and unmanned devices. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 5672–5697. [\[CrossRef\]](#)
- Song, J.; Zhou, X.; Xu, C.; Xu, S.; Shu, J. Training-free automatic instance segmentation of girder bridge point cloud via large model fusion with reverse entity modelling verification. *Autom. Constr.* **2025**, *179*, 106484. [\[CrossRef\]](#)
- Yang, Z.; Shu, J.; Jiang, J.; Han, W.; Wang, Y.; Zhao, L.; Bai, Y. Automated path-planning strategy for robotic inspection of underground utilities based on building information model. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 5554–5575. [\[CrossRef\]](#)
- Tang, Y.; Wang, Y.; Gao, X.; Xie, H.; Tan, H. Task and Motion Planning for Mobile Tracking Robot in Automatic Inspection of Large-Scale Freeform Surface Parts. *IEEE Trans. Ind. Electron.* **2026**, *14*. [\[CrossRef\]](#)
- Sun, Z.Y.; Guo, Y.T.; Ge, K.; Hou, C.; Hu, Z.Z. Deep learning-based realtime multiload response prediction and inverse analysis of offshore bridges. *Eng. Struct.* **2026**, *353*, 122330. [\[CrossRef\]](#)
- Yu, Y.; Liu, H.; Fu, Y.; Jia, W.; Yu, J.; Yan, Z. Embedding Pose Information for Multiview Vehicle Model Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5467–5480. [\[CrossRef\]](#)
- Cho, H.S.; Latif, K.; Sharafat, A.; Seo, J. Multi-Modal Excavator Activity Recognition Using Two-Stream CNN-LSTM with RGB and Point Cloud Inputs. *Appl. Sci.* **2025**, *15*, 8505. [\[CrossRef\]](#)
- Assadzadeh, A.; Arashpour, M.; Brilakis, I.; Ngo, T.; Konstantinou, E. Vision-based excavator pose estimation using synthetically generated datasets with domain randomization. *Autom. Constr.* **2022**, *134*, 104089. [\[CrossRef\]](#)
- Chen, F.; Yu, Y.; Yi, J.; Zhang, T.; Zhao, J.; Jia, W.; Yu, J. MCLL-Diff: Multiconditional Low-Light Image Enhancement Based on Diffusion Probabilistic Models. *IEEE Sens. J.* **2025**, *25*, 9912–9924. [\[CrossRef\]](#)
- Wang, M.; Zhou, D.; Chen, M. Hybrid variable monitoring: An unsupervised process monitoring framework with binary and continuous variables. *Automatica* **2023**, *147*, 110670. [\[CrossRef\]](#)
- Yan, D.; Li, R.; Xiong, W.; Huang, X. Sensor selection strategy and multimodal layout optimization for autonomous Vehicles: A review. *Measurement* **2026**, *274*, 121225. [\[CrossRef\]](#)
- Molaei, A.; Kolu, A.; Lahtinen, K.; Geimer, M. Automatic recognition of excavator working cycles using supervised learning and motion data obtained from inertial measurement units (IMUs). *Constr. Robot.* **2024**, *8*, 14. [\[CrossRef\]](#)
- Bae, J.; Kim, K.; Hong, D. Automatic Identification of Excavator Activities Using Joystick Signals. *Int. J. Precis. Eng. Manuf.* **2019**, *20*, 2101–2107. [\[CrossRef\]](#)
- Molaei, A.; Kolu, A.; Lahtinen, K.; Geimer, M. Automatic estimation of excavator actual and relative cycle times in loading operations. *Autom. Constr.* **2023**, *156*, 105080. [\[CrossRef\]](#)
- Weng, S.; Zhang, Z.; Gao, K.; Hu, B.; Zhang, J.; Zhu, H. Revisiting the sensing mechanism of overlapping graphene based flexible strain sensors. *Compos. Sci. Technol.* **2026**, *277*, 111542. [\[CrossRef\]](#)
- Wen, J.; Li, B.; Ren, L.; Wang, K.; Cao, Y.; Ren, L. Biomimetic additive manufacturing tactile sensing systems: Mechanisms, materials, techniques, and prospects. *Mater. Today* **2026**, *92*, 711–750. [\[CrossRef\]](#)
- Zhang, Y.; Zhao, X.; Xu, R.; Guo, R.; Wei, X. A dual-path time-frequency integration architecture applied to remaining useful life prediction of air turbine starter bearing. *Meas. Sci. Technol.* **2026**, *37*, 086201. [\[CrossRef\]](#)
- Wang, Z.; Oates, T. Imaging Time-Series to Improve Classification and Imputation. *arXiv* **2015**, arXiv:1506.00327. [\[CrossRef\]](#)
- Wan, A.; Tong, X.; Al-Bukhaiti, K.; Zhou, Z.; Su, Y.; Cheng, X.; Ji, X. Vibration prediction for abnormal elevator door system faults based on attention mechanism and neural networks with time-frequency domain features. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2025**, *239*, 7358–7372. [\[CrossRef\]](#)

28. Xie, Z.; Mo, C.; Jia, B. CDFAF: A co-evolutionary decoupled attention framework for explainable weak thermal fault diagnosis of marine diesel engines. *Sci. Rep.* **2026**, *16*, 17459. [[CrossRef](#)] [[PubMed](#)]
29. Wang, J.; Zhang, R.; Gu, Q.; Skitmore, M.; Chileshe, N.; Qu, Z.; Wang, Z.; Wang, X.; Liu, H. Time Series Forecasting in Construction Management: A Scientometric Analysis, Qualitative Review, and Future Research. *Buildings* **2026**, *16*, 3496. [[CrossRef](#)]
30. Kim, J.; Chi, S.; Seo, J. Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks. *Autom. Constr.* **2018**, *87*, 297–308. [[CrossRef](#)]
31. Mahmood, B.; Han, S.U.; Seo, J. Implementation experiments on convolutional neural network training using synthetic images for 3D pose estimation of an excavator on real images. *Autom. Constr.* **2022**, *133*, 103996. [[CrossRef](#)]
32. Pham, H.T.T.L.; Han, S.U. Generating realistic training images from synthetic data for excavator pose estimation. *Autom. Constr.* **2024**, *167*, 105718. [[CrossRef](#)]
33. Liu, G.; Wang, Q.; Wang, T.; Li, B.; Xi, X. Vision-based excavator pose estimation for automatic control. *Autom. Constr.* **2024**, *157*, 105162. [[CrossRef](#)]
34. Soltani, M.M.; Zhu, Z.; Hammad, A. Skeleton estimation of excavator by detecting its parts. *Autom. Constr.* **2017**, *82*, 1–15. [[CrossRef](#)]
35. Shin, Y.; Seo, S.; Koo, C. Deep learning-based automated method for enhancing excavator activity recognition in far-field construction site surveillance videos. *Autom. Constr.* **2025**, *173*, 106099. [[CrossRef](#)]
36. Liu, Y.; Jia, T.; Wei, J.; Ma, B.; Wang, H.; Chen, D. GAA-TSO: Geometry-Aware-Assisted Depth Completion for Transparent and Specular Objects. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 7747–7763. [[CrossRef](#)]
37. Liu, L.; Cai, C.; Shen, S.; Liang, J.; Ouyang, W.; Ye, T.; Mao, J.; Duan, H.; Yao, J.; Zhang, X.; et al. MoA-VR: A Mixture-of-Agents System Toward All-in-One Video Restoration. *IEEE J. Sel. Top. Signal Process.* **2025**, *19*, 1822–1837. [[CrossRef](#)]
38. Peng, X.; Zhang, Y.; Zhang, X.; Wang, J.; Bai, S.; Cao, Y.; Chen, H.; Li, T. FCS-edNET: Exploring Magnetic Particle Imaging Deblurring With Neural Network. *IEEE Trans. Image Process.* **2026**, *35*, 480–494. [[CrossRef](#)] [[PubMed](#)]
39. Wu, X.; He, Z.; Jiang, G.; Yu, M.; Song, Y.; Luo, T. No-Reference Point Cloud Quality Assessment Through Structure Sampling and Clustering Based on Graph. *IEEE Trans. Broadcast.* **2025**, *71*, 307–322. [[CrossRef](#)]
40. He, Z.; Liang, Q.; Jiang, G.; Yu, M.; Chen, Y.; Luo, T.; Zhou, W. Local and Global Structure-Guided No-Reference Point Cloud Quality Assessment. *IEEE Trans. Multimed.* **2025**, *27*, 9252–9266. [[CrossRef](#)]
41. Zhang, Q.; Wang, J.; Shen, Y.; Zhang, B.; Feng, C.; Pan, J. Privilege-guided knowledge distillation for edge deployment in excavator activity recognition. *Autom. Constr.* **2024**, *166*, 105688. [[CrossRef](#)]
42. Shen, Y.; Wang, J.; Mo, S.; Gu, X. Data augmentation aided excavator activity recognition using deep convolutional conditional generative adversarial networks. *Adv. Eng. Inform.* **2024**, *62*, 102785. [[CrossRef](#)]
43. Rashid, K.M.; Louis, J. Times-series data augmentation and deep learning for construction equipment activity recognition. *Adv. Eng. Inform.* **2019**, *42*, 100944. [[CrossRef](#)]
44. Rashid, K.M.; Louis, J. Automated Activity Identification for Construction Equipment Using Motion Data From Articulated Members. *Front. Built Environ.* **2020**, *5*, 144. [[CrossRef](#)]
45. Akhavian, R.; Behzadan, A.H. Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. *Adv. Eng. Inform.* **2015**, *29*, 867–877. [[CrossRef](#)]
46. Song, H.; Li, G.; Xiong, X.; Li, M.; Qin, Q.; Mitrouchev, P. A novel data fusion based intelligent identification approach for working cycle stages of hydraulic excavators. *ISA Trans.* **2024**, *148*, 78–91. [[CrossRef](#)] [[PubMed](#)]
47. Bai, J.; Zhu, W.; Liu, S.; Ye, C.; Zheng, P.; Wang, X. A Temporal Convolutional Network-Bidirectional Long Short-Term Memory (TCN-BiLSTM) Prediction Model for Temporal Faults in Industrial Equipment. *Appl. Sci.* **2025**, *15*, 1702. [[CrossRef](#)]
48. Wan, A.; Zhu, Z.; AL-Bukhaiti, K.; Cheng, X.; Jiang, J.; Ji, X.; Wang, J.; Shan, T. Real-time aero-engine fault diagnosis using 5G edge computing and deep learning. *Measurement* **2026**, *257*, 118784. [[CrossRef](#)]
49. Hong, S.; Yoon, J.; Ham, Y.; Lee, B.; Kim, H. Monitoring safety behaviors of scaffolding workers using Gramian angular field convolution neural network based on IMU sensing data. *Autom. Constr.* **2023**, *148*, 104748. [[CrossRef](#)]
50. Kim, J.; Chi, S.; Ahn, C.R. Hybrid kinematic-visual sensing approach for activity recognition of construction equipment. *J. Build. Eng.* **2021**, *44*, 102709. [[CrossRef](#)]
51. Deng, T.; Sharafat, A.; Lee, S.; Seo, J. Automatic vision-based volume estimation of dump loading for real-time earthwork productivity assessment. *Expert Syst. Appl.* **2026**, *303*, 130657. [[CrossRef](#)]
52. Cao, Y.; Zhang, R.; Qu, Z.; Skitmore, M.; Ma, X.; Wang, J. Forecasting Fatal Construction Accidents Using an STL-BiGRU Hybrid Framework: A Multi-Scale Time Series Approach. *Buildings* **2026**, *16*, 1539. [[CrossRef](#)]
53. Liang, D.; Hu, H.; Wu, Y. Geometric design and dynamic characteristics analysis of a novel herringbone planetary gear. *Proc. Inst. Mech. Eng. Part K J. Multi-Body Dyn.* **2025**, *239*, 234–253. [[CrossRef](#)]
54. Shen, Z.; He, Y.; Du, X.; Yu, J.; Wang, H.; Wang, Y. YCANet: Target Detection for Complex Traffic Scenes Based on Camera-LiDAR Fusion. *IEEE Sens. J.* **2024**, *24*, 8379–8389. [[CrossRef](#)]

55. Wang, Q.; Guo, H.; Yang, C.; He, Y.; Chen, X.; Xu, B. A study on the enhancement method for seam extraction in teachless welding robots based on a multichannel feature fusion network. *Int. J. Adv. Manuf. Technol.* **2025**, *141*, 647–660. [[CrossRef](#)]
56. Ma, B.; Jia, T.; Wang, H.; Chen, D. Meta-TIP: An Unsupervised End-to-End Fusion Network for Multi-Dataset Style-Adaptive Threat Image Projection. *IEEE Trans. Image Process.* **2025**, *34*, 8317–8331. [[CrossRef](#)] [[PubMed](#)]
57. Cheng, Y.; Lu, M.; Gai, X.; Guan, R.; Zhou, S.; Xue, J. Research on multi-signal milling tool wear prediction method based on GAF-ResNext. *Robot. Comput.-Integr. Manuf.* **2024**, *85*, 102634. [[CrossRef](#)]
58. Guo, R.; Yi, J.; Luo, X. An efficient classification and error correction knowledge distillation framework for remaining useful life prediction of bearings in air turbine starter. *Neurocomputing* **2025**, *657*, 131611. [[CrossRef](#)]
59. Wan, A.; Zhang, F.; AL-Bukhaiti, K.; Cheng, X.; Ji, X. Robust bearing remaining useful life prediction using a hybrid deep learning framework across operating conditions. *J. Braz. Soc. Mech. Sci. Eng.* **2026**, *48*, 230. [[CrossRef](#)]
60. Luo, Y.; Zhu, M.; Chen, T.; Zheng, Z. Remaining useful life prediction for stratospheric airships based on a channel and temporal attention network. *Commun. Nonlinear Sci. Numer. Simul.* **2025**, *143*, 108634. [[CrossRef](#)]
61. Assadzadeh, A.; Arashpour, M.; Li, H.; Hosseini, R.; Elghaish, F.; Baduge, S. Excavator 3D pose estimation using deep learning and hybrid datasets. *Adv. Eng. Inform.* **2023**, *55*, 101875. [[CrossRef](#)]
62. Zhao, X.; Jia, K.; Zhang, Y.; Sun, Z.; Feng, J.; Liu, W. GAF-Based Multimodal Time Series Data Fusion for Water Depth Monitoring in Headwater Streams. In *Proceedings—2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*; IEEE: New York, NY, USA, 2024; pp. 174–181. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.