AI-Based Multimodal Anomaly Detection for Industrial Machine Operations

Qiaoyun Zhang[1], Hsiang-Chuan Chang[3], Chia-Ling Ho[4], Huan-Chao Keh[2], Diptendu Sinha Roy[5]*

[1] School of Artificial Intelligence, Chuzhou University, China
[2] Department of Computer Science and Information Engineering, Tamkang University, Taiwan
[3] Department of Transportation Management, Tamkang University, Taiwan
[4] General Education Center, National Taipei University of Nursing and Health Sciences, Taiwan
[5] Department of Computer Science and Engineering, National Institute of Technology, India
zqyun@chzu.edu.cn, 149190@o365.tku.edu.tw, chialingho@ntunhs.edu.tw,
hckeh@mail.tku.edu.tw, diptendu.sr@nitm.ac.in

Abstract

In the manufacturing process involving grinding wheels, challenges in fine-tuning grinding machines are typically addressed by craftsmen through subjective observations of sparks and sounds. However, most current anomaly detection methods mainly aim at a single modality, whereas existing multimodal methods cannot effectively cope with a common issue. To address this, this paper introduces an innovative mechanism, AI-Based Multimodal Anomaly Detection (AMAD), designed to optimize the efficiency and accuracy of grinding wheel production lines. The proposed AMAD includes data preprocessing and multimodal anomaly detection, accurately identifying anomalies in grinding wheel operation videos. In the data preprocessing phase, the proposed AMAD utilizes Mel Frequency Cepstral Coefficients (MFCC) and AutoEncoder for audio processing and segmentation for video processing. In the multimodal anomaly detection phase, the proposed AMAD employs Convolutional Neural Networks (CNN) for audio analysis and Convolutional Long Short-Term Memory (ConvLSTM) for video analysis. By combining both audio and video modalities, the proposed AMAD effectively predicts whether the input video represents normal or abnormal grinding wheel operations. This multimodal approach not only improves the accuracy of anomaly detection but also enhances the robustness of the system. Simulation results demonstrate that the proposed AMAD significantly improves performance in anomaly detection in terms of precision, recall, and F1-Score.

Keywords: MFCC, ConvLSTM, CNN, Anomaly Detection

1 Introduction

Anomaly detection in industrial processes is crucial for maintaining operational efficiency, safety, and cost-effectiveness. The detection of abnormal events in manufacturing systems, such as grinding wheel operations, can prevent significant downtime and equipment damage, ensuring the seamless flow of production. Traditionally, anomaly detection has relied heavily on manual inspections and simple threshold-based methods, which often fail to capture complex, nuanced abnormalities within industrial environments.

In recent years, the advent of machine learning [1-3] and deep learning [4-7] techniques have provided promising alternatives for automated anomaly detection. These methods have demonstrated remarkable capabilities in identifying irregularities across various domains by analyzing large datasets, extracting relevant features, and learning intricate patterns that might indicate abnormal behavior. However, many existing approaches [1-7] are limited by their reliance on unimodal data, which may not fully capture the multi-faceted nature of industrial processes, especially in environments where both audio and visual signals are informative.

To address these limitations, this paper introduces an innovative anomaly detection mechanism for grinding wheel operations, called, AMAD. AMAD leverages multimodal data by integrating both audio and video information to enhance the detection accuracy of abnormal operations. It utilizes MFCC [8] and AutoEncoder [9] for audio preprocessing and segmentation techniques for video preprocessing. It employs CNN [10] for audio analysis and ConvLSTM [11-12] for video analysis, combining both modalities to accurately classify grinding wheel operations as normal or abnormal. This multimodal fusion provides a robust solution, significantly outperforming traditional methods. The main contributions of the paper are summarized as follows:

(1) Utilizing MFC and AutoEncoder to extract features of audio: the proposed AMAD utilizes MFCC for audio feature extraction and employs AutoEncoder for feature learning and dimensionality reduction. This advanced audio processing method effectively captures key information in the audio data, providing a reliable foundation for subsequent anomaly detection.

(2) Integrating both audio and video modalities to improve the accuracy of anomaly detection: the proposed AMAD combines both audio and video modalities, using CNN for audio analysis and ConvLSTM

*Corresponding Author: Huan-Chao Keh; E-mail: hckeh@mail.tku.edu.tw
DOI: https://doi.org/10.70003/160792642025032602010

for video analysis. This comprehensive monitoring and anomaly detection of grinding wheel operations through multimodal data fusion not only improves the accuracy of anomaly detection but also enhances the robustness, significantly increasing the efficiency and precision of the grinding wheel production line.

(3) Dynamically adjusting weights for joint loss functions: the proposed AMAD utilizes an adaptive training strategy that involves joint optimization of audio and video loss functions with dynamic adjustment of weight parameters. This approach not only accelerates the model's convergence but also strengthens its adaptability and generalization ability to various anomalous situations.

The remainder of the paper is organized as follows. Section 2 reviews and contrasts previous relevant work, while Section 3 outlines the assumptions and problem descriptions. Section 4 details the proposed AMAD mechanism, and Section 5 focuses on the performance study. Finally, Section 6 concludes with a summary and discussion of future work.

2 Related Work

Recent research in anomaly detection has focused on ensuring system reliability, with solutions generally categorized into single-modality and multimodal detection.

2.1 Single-Modality Anomaly Detection

Single-modality anomaly detection primarily extracts features from a single data source, such as audio or video, to identify anomalies [1-7]. These approaches are divided into traditional machine learning methods [1-3, 18-19] and deep learning techniques [4-7]. Machine learning-based methods, such as Scudo et al. [1] for audio anomaly detection and Wu et al. [3] for unsupervised industrial audio detection, often struggle with complex, high-dimensional data and fail to capture spatial and temporal dependencies.

To overcome these limitations, deep learning methods have been used to automatically extract spatial features via CNNs and model temporal dependencies with LSTMs. For example, Jagadeeshwar et al. [4] applied CNN-based emotion recognition to audio, while Zou et al. [5] used few-shot learning for mechanical anomaly detection. Other notable contributions include Kulkarni et al. [6] for respiratory anomaly detection and Wang et al. [7] for machine sound detection. Despite the notable success of these single-modality anomaly detection methods, they still face challenges when dealing with complex, multi-source data environments. Therefore, the proposed AMAD explores multimodal anomaly detection methods that integrate audio and visual data, aiming to achieve better performance across different application scenarios.

2.2 Multimodal Anomaly Detection

While single-modal anomaly detection methods have been proven effective, Lee et al. [13] noted that relying on a single data source may miss certain anomalies detectable through multimodal analysis. Multimodal approaches address this by combining data from multiple sources for a more comprehensive understanding. For example, Wang et al. [14] used multimodal data in microservice systems to enhance detection accuracy, though distinguishing between normal hard samples and anomalies remained challenging. Liu et al. [15] tackled this issue with contrastive learning and adversarial training to separate hard samples from anomalies, but these methods require extensive labeled data, limiting their scalability.

In audio-visual applications, Feng et al. [16] introduced a self-supervised video forensics method leveraging audio-visual anomaly detection, showcasing the potential of multimedia data. Similarly, Gao et al. [17] utilized audio-visual representation learning for crowd anomaly detection, offering a more holistic solution.

Despite advancements in multimodal approaches, challenges remain due to the complexity and heterogeneity of multimodal data, including issues with data alignment and inter-modal fusion. To overcome these challenges, the proposed AMAD employs MFCC and AutoEncoder for audio preprocessing and segmentation techniques for video alignment. This preprocessing strategy facilitates seamless multimodal fusion, ultimately enhancing the accuracy and reliability of anomaly detection.

Table 1 compares the proposed AMAD with related work. The 'Method' column indicates the type of mechanism used, while the 'Modality' column specifies whether the mechanism handles single or multimodal data. The 'Spatial', 'Temporal', and 'Audio Feature Generation' columns denote whether the mechanism accounts for these aspects in its analysis. Compared to related work, the proposed AMAD leverages deep learning to process multimodal data effectively, integrating spatial and temporal features and employing advanced audio feature processing techniques for superior anomaly detection.

Table 1. The comparisons of the proposed AMAD and related work

Mechanism	Method	Modality	Spatial	Temporal	Audio feature generation
[1-3]	Machine learning	Single	✕	✕	✕
[19]	Machine learning	Single	○	✕	✕
[4-5]	Machine/ Deep learning	Single	○	✕	○
[6]	Machine/ Deep learning	Single	✕	○	✕
[7]	Machine/ Deep learning	Single	○	○	✕
[16]	Deep learning	Multimodal	○	○	✕
[17]	Deep learning	Multimodal	○	○	✕
[18]	Deep learning	Multimodal	○	○	✕
AMAD (Ours)	Deep learning	Multimodal	○	○	○

3 Notations, Assumptions and Problem Descriptions

This section introduces the notations, assumptions, problem descriptions, and an objective of this paper.

3.1 Notations and Assumptions

Assume that there is a set of n operational videos of grinding wheels, denoted by $\mathbb{V} = \{V_1, V_2, \ldots, V_n\}$. Each video $V_i \in \mathbb{V}$ is divided into two tracks: the image track and the audio track, denoted by I_i and A_i, respectively. Let $\mathbb{I} = \{I_1, I_2, \ldots, I_n\}$ denote a set of images, which capture visual information about the grinding process. These images are extracted at regular intervals to provide a sequence of frames that can be analyzed to detect visual anomalies in the grinding operation.

Similarly, let $\mathbb{A} = \{A_1, A_2, \ldots, A_n\}$ denote a set of n audios, which records the sound associated with the grinding process. By analyzing both the visual and auditory data, it becomes possible to accurately identify anomalies in the grinding wheel operations.

3.2 Problem Descriptions

This paper leverages a confusion matrix to assess the effectiveness of anomaly detection in grinding wheel operations. Let y_i^I be a Boolean variable representing the ground truth of whether the anomaly can be detected from the track I_i. Let $\hat{y}_i^I$ denote the prediction of the image track I_i by applying mechanism M. The values of y_i^I and $\hat{y}_i^I$ can be derived by Eqs. (1) and (2), respectively.

$$y_i^I = \begin{cases} 1, I_i \text{ represents a normal grinding wheel operation}, \\ 0, \quad \text{otherwise.} \end{cases} \tag{1}$$

$$\hat{y}_i^I = \begin{cases} 1, I_i \text{ is predicted as a normal grinding wheel operation}, \\ 0, \quad \text{otherwise.} \end{cases} \tag{2}$$

Let y_i^A denote the ground truth of whether the anomaly can be detected from the audio track A_i. Let $\hat{y}_i^A$ denote the prediction of the audio track A_i. The values of y_i^A and $\hat{y}_i^A$ can be derived by Eqs. (3) and (4), respectively.

$$y_i^A = \begin{cases} 1, \text{ if } A_i \text{ is a normal sound}, \\ 0, \quad \text{otherwise.} \end{cases} \tag{3}$$

$$\hat{y}_i^A = \begin{cases} 1, \text{ if } A_i \text{ is predicted as a normal sound}, \\ 0, \quad \text{otherwise.} \end{cases} \tag{4}$$

Let TP_i^M denote the true positive of the result predicted by the mechanism M. That is, for $V_i \in \mathbb{V}$, the predictions of $\hat{y}_i^I$ and $\hat{y}_i^A$ are both correct. The value of TP_i^M can be derived by Eq. (5).

$$TP_i^M = \begin{cases} 1, & \text{if } (\hat{y}_i^I \cdot y_i^I) + (\hat{y}_i^A \cdot y_i^A) \geq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Similarly, let FP_i^M and FN_i^M denote false positive and false negative of the prediction result for the i-th video V_i, respectively. The values of FP_i^M and FN_i^M can be calculated by Eqs. (6) and (7), respectively.

$$FP_i^M = \begin{cases} 1, & \text{if } (\hat{y}_i^I - y_i^I) + (\hat{y}_i^A - y_i^A) \geq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$$FN_i^M = \begin{cases} 1, & \text{if } (y_i^I - \hat{y}_i^I) + (y_i^A - \hat{y}_i^A) \geq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Let TP^M, FP^M, and FN^M denote true positive, false positive, and false negative by applying mechanism M for anomaly detection of $\mathbb{V}$, respectively. The values of TP^M, FP^M, and FN^M can be calculated by Eqs. (8) to (10), respectively.

$$TP^M = \sum_{i=1}^{n} TP_i^M, \tag{8}$$

$$FP^M = \sum_{i=1}^{n} FP_i^M, \tag{9}$$

$$FN^M = \sum_{i=1}^{n} FN_i^M. \tag{10}$$

The precision and recall of all $V_i \in \mathbb{V}$, denoted by $\mathcal{P}^M$ and $\mathcal{R}^M$ respectively, can be calculated as follows:

$$\mathcal{P}^M = \frac{TP^M}{TP^M + FP^M} = \sum_{i=1}^{n} TP_i^M \Big/ \left(\sum_{i=1}^{n} TP_i^M + \sum_{i=1}^{n} FP_i^M \right), \tag{11}$$

$$\mathcal{R}^M = \frac{TP^M}{TP^M + FN^M} = \sum_{i=1}^{n} TP_i^M \Big/ \left(\sum_{i=1}^{n} TP_i^M + \sum_{i=1}^{n} FN_i^M \right). \tag{12}$$

Let $F1^M$ denote $F1$-score of all $V_i \in \mathbb{V}$. The value of $F1^M$ can be denoted by Exp. (13).

$$F1^M = \frac{2 * \mathcal{P}^M * \mathcal{R}^M}{\mathcal{P}^M + \mathcal{R}^M}. \tag{13}$$

3.3 Objective

Let Ω denote a set of all possible mechanisms for abnormal detection in grinding wheel operations. The primary objective of the proposed AMAD is to find the best mechanism M_{best} which satisfies the Exp. (14).

Objective Function:

$$M_{best} = arg \max_{M \in \Omega} F1^{M} . \quad (14)$$

4 The Proposed SMART Mechanism

This paper introduces an innovative anomaly detection mechanism for grinding wheel operations, called AMAD. AMAD accurately identify abnormal operations in grinding wheel operation videos. As illustrated in Figure 1, the proposed AMAD consists of two key phases: data preprocessing and multimodal anomaly detection.

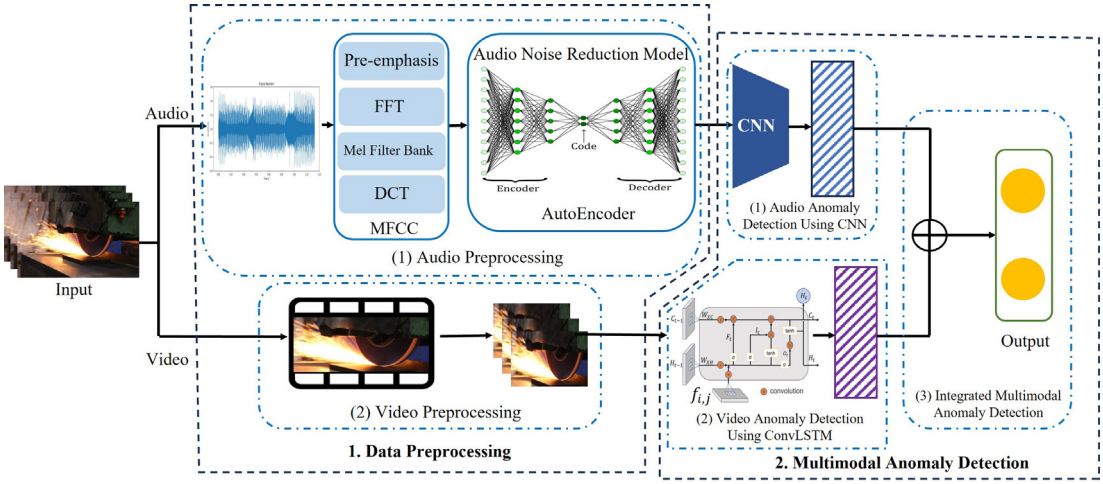


Figure 1. The architecture of the AMAD mechanism

In the data preprocessing phase, MFCC and an AutoEncoder are used for audio processing, while segmentation handles video processing, ensuring the data is well-prepared for analysis. In the multimodal anomaly detection phase, the proposed AMAD employs CNN for audio analysis and ConvLSTM for video analysis. By combining these two modalities, AMAD effectively predicts whether the input video represents normal or abnormal grinding wheel operations.

4.1 Data Preprocessing

Given a set of n videos $\mathbb{V} = \{V_1, V_2, \ldots, V_n\}$, this phase aims to split each grinding wheel operation video $V_i \in \mathbb{V}$ into the audio track A_i and image track I_i , followed by preprocessing of both tracks. This phase consists of two tasks: Audio Preprocessing and Video Preprocessing. The Audio Preprocessing utilizes MFCC, and AutoEncoder to process the audio data while the Video Preprocessing segments each video I_i into fixed-period clips, preparing the visual data for further analysis.

4.1.1 The Task of Audio Preprocessing and Feature Extraction

This task aims to apply MFCC extraction followed by AutoEncoder processing to each audio sample. It consists of two steps: MFCC extraction and AutoEncoder processing.

(1) *MFCC Extraction Step*

Let $\mathcal{A}_i$ denote the output of MFCC extraction for A_i. The value of $\mathcal{A}_i$ can be calculated by Eq. (15).

$$\mathcal{A}_i = D_{P\times M} \cdot \log\left(H_{M\times K} \cdot \left| F_{K\times L}\left(w_L \odot f_{N,L}\left(P_\alpha\left(A_i \right) \right) \right) \right|^2 \right), \quad (15)$$

where $P_\alpha(\cdot)$ denotes pre-emphasis operation with coefficient α. The notation $f_{N,L}$ denotes framing operation, segmenting the A_i into N overlapping frames of length L. The w_L is hamming window function of length L. The $F_{K\times L}(\cdot)$ denotes Discrete Fourier Transform, converting L time-domain samples to K frequency-domain coefficients. Finally, the $H_{M\times K}$ denotes Mel filter bank matrix, containing M filters and the $D_{P\times M}$ denotes Discrete Cosine Transform, used to extract P MFCC coefficients.

(2) *AutoEncoder Processing Step*

Following MFCC extraction, each feature $\mathcal{A}_i$ is further processed through an AutoEncoder model for dimensionality reduction and feature learning. Let $E(\cdot)$ and $D(\cdot)$ denote the encoder and decoder functions of the AutoEncoder model, respectively. Let ε_i denote the output of the encoder function, defined as:

$$\mathcal{E}_i = E\left(\mathcal{A}_i\right) = \sigma\left(W_e \cdot \mathcal{A}_i + b_e\right), \quad (16)$$

where $\sigma(\cdot)$ is a non-linear activation function, W_e is the encoding weight matrix, and b_e is the encoding bias vector. Let $\mathcal{D}_i$ denote the output of the decoder function. That is

$$\mathcal{D}_i = D\left(\mathcal{E}_i\right) = \sigma\left(W_d \cdot \mathcal{E}_i + b_d\right), \quad (17)$$

where W_d is the decoding weight matrix, and b_d is the decoding bias vector. Let $\mathcal{L}_A(\cdot)$ denote the loss function of the AutoEncoder, which uses mean squared error. The value of $\mathcal{L}_A(\cdot)$ can be calculated by Eq. (18).

$$\mathcal{L}_A\left(\mathcal{D}_i, \mathcal{A}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{D}_i - \mathcal{A}_i\right)^2, \tag{18}$$

where n denotes the number of audio tracks. This loss function guides the training of the AutoEncoder, ensuring that the encoder effectively extracts meaningful features. The output $\mathcal{D}_i$ serves as the input for subsequent audio anomaly detection task.

4.1.2 The Task of Video Preprocessing

This task outlines the preprocessing steps applied to each video sample, focusing on segmentation processing and frame extraction. These procedures ensure uniformity across all video samples in the dataset. Recall that $\mathbb{I} = \{I_1, I_2, \ldots, I_n\}$ denotes a set of video samples. Each video $I_i \in \mathbb{I}$ is processed to a standardized length L_v. The value of L_v is determined by the longest video in the dataset, as defined in Eq. (19).

$$L_v = \max_{\forall I_i \in \mathbb{I}} L\left(I_i\right), \tag{19}$$

where $L(\cdot)$ is a function that returns the length of a given video.

For video $I_i \in \mathbb{I}$ with lengths less than L_v, padding is applied to reach the standardized length. Let $\mathcal{V}_i$ denote the padded version of I_i. The padded video $\mathcal{V}_i$ can be expressed by Eq. (20).

$$\mathcal{V}_i = \left[I_i,\ Pad\left(L_v - L\left(I_i\right)\right)\right], \tag{20}$$

where $Pad(\cdot)$ is a padding function that generates the required number of frames to achieve the standardized length.

For each standardized video $\mathcal{V}_i$, frames are extracted at a constant rate of N_v per second. This process yields a set of frames, denoted by $F_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,Lv\cdot Nv}\}$, where each frame $f_{i,j} \in F_i$ is captured at $\frac{j}{N_v}$ seconds into the video $\mathcal{V}_i$. That is

$$f_{i,j} = \mathcal{V}_i\left(\frac{j}{N_v}\right). \tag{21}$$

This can ensure uniform sampling across all videos in the dataset, facilitating consistent feature extraction and subsequent analysis.

4.2 Multimodal Anomaly Detection

This phase aims to utilize CNN and ConvLSTM models for detecting anomalies in audio and video anomaly, respectively. It is divided into three tasks: audio anomaly detection using CNN, video anomaly detection using ConvLSTM, and combined multimodal anomaly detection.

4.2.1 Task of Audio Anomaly Detection Using CNN

This task aims to detect audio anomalies using CNN. Recall that $\mathcal{D}_i$ is the preprocessed audio features for the i-th sample. The CNN model for audio can be represented by the function $\mathcal{F}_{audio}$. That is

$$\mathcal{F}_{audio}\left(\mathcal{D}_i\right) = FC\left(Conv_k\left(\ldots Pool\left(Conv_1\left(\mathcal{D}_i\right)\right)\right)\right), \tag{22}$$

where $Conv_k(\cdot)$ denotes the k-th convolutional layer, $Pool(\cdot)$ denotes the pooling layer, and $FC(\cdot)$ denotes the fully connected layer. Recall that $y_i^A \in \mathcal{R}^2$ and $\hat{y}_i^A \in \mathcal{R}^2$ denote the ground truth and prediction for the audio track A_i, respectively. This task is mainly for binary classification, including normal and abnormal classes. Let $\mathcal{L}_{audio}$ denote the loss function for the audio anomaly detection. The $\mathcal{L}_{audio}$ is represented by

$$\mathcal{L}_{audio}\left(y_i^A, \hat{y}_i^A\right) = -\frac{1}{n}\sum_{i=1}^{n} y_i^A \log\left(\hat{y}_i^A\right), \tag{23}$$

where n is the number of audio samples.

4.2.2 Task of Video Anomaly Detection Using ConvLSTM

This task aims to utilize ConvLSTM to detect video anomalies. Recall that $F_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,Lv\cdot Nv}\}$ represents the preprocessed video frames for the i-th video I_i. Let $\mathcal{F}_{video}$ denote the function of the ConvLSTM model for video. That is

$$\begin{aligned}&\mathcal{F}_{video}\left(F_i\right)\\ &= FC\left(ConvLSTM_k\left(\ldots ConvLSTM_2\left(ConvLSTM_1\left(F_i\right)\right)\right)\right)\end{aligned}, \tag{24}$$

where $ConvLSTM_k(\cdot)$ denotes the k-th ConvLSTM layer. Recall that $y_i^I \in \mathcal{R}^2$ and $\hat{y}_i^I \in \mathcal{R}^2$ denote the ground truth and prediction of the video track I_i, respectively. Let $\mathcal{L}_{video}$ denote the loss function for the video anomaly detection. That is

$$\mathcal{L}_{video}\left(y_i^I, \hat{y}_i^I\right) = -\frac{1}{n}\sum_{i=1}^{n} y_i^I \log\left(\hat{y}_i^I\right), \tag{25}$$

where n is the number of video samples.

4.2.3 Task of Integrated Multimodal Anomaly Detection

To improve the overall detection performance, this task employs integrated multimodal anomaly detection by fusing the features or detection results from the audio and video models. The final outputs of the audio and video models are combined as follows.

$$\mathbb{C}_i^c = \sigma\left(W_A \cdot \mathcal{F}_{audio}\left(\mathcal{D}_i\right) + W_I \cdot \mathcal{F}_{video}\left(F_i\right)\right), \tag{26}$$

where W_A and W_I are the weights for the audio and video outputs, respectively, and σ is the activation function.

Let $y_i^c \in \mathcal{R}^2$ and $\hat{y}_i^c \in \mathcal{R}^2$ denote the ground truth and prediction of the integrated audio and video, respectively. The predicted label of $\hat{y}_i^c$ can be derived from Eq. (27).

$$\hat{y}_i^c = softmax\left(W_B \cdot \mathbb{C}_i^c + b_B\right), \quad \textbf{(27)}$$

where W_B is the weight for the combined outputs, and b_B is the bias. Let $\mathcal{L}_{integrated}$ denote the loss function for the combined modal. The value of $\mathcal{L}_{combined}$ can be derived from Eq. (28).

$$\mathcal{L}_{integrated}\left(y_i^c, \hat{y}_i^c\right) = -\frac{1}{n}\sum_{i=1}^{n} y_i^c \log\left(\hat{y}_i^c\right). \quad \textbf{(28)}$$

The total loss function $\mathcal{L}_{total}$ for training the multimodal anomaly detection system is a weighted sum of the individual loss functions:

$$\mathcal{L}_{total} = \beta \cdot \mathcal{L}_{audio} + \gamma \cdot \mathcal{L}_{video} + \left(1-\beta-\gamma\right) \cdot \mathcal{L}_{combined}, \quad \textbf{(29)}$$

where β, and γ are weight coefficients used to balance the contributions of the audio and video losses. The objective of optimization is to minimize the total loss function. That is

$$\min_{\forall i} \mathcal{L}_{total}. \quad \textbf{(30)}$$

Table 2 describes the algorithm of the proposed AMAD, consisting of two phases: data preprocessing and multimodal anomaly detection. In the data preprocessing phase, input videos are separated into audio and image tracks. Audio data is feature-extracted using MFCC and AutoEncoder, while video data is normalized, and frames are extracted. In the multimodal anomaly detection phase, CNN processes the audio data, and ConvLSTM processes the video data. Theie results are integrated for comprehensive detection. The algorithm optimizes model parameters by minimizing the total loss function, outputting trained audio and video anomaly detection models ($\mathcal{F}_{video}$ and $\mathcal{F}_{audio}$). The proposed AMAD effectively combines audio and video information to improve the accuracy and robustness of anomaly detection.

Table 2. The algorithm of the proposed AMAD mechanism

Input:	$\mathbb{V} = \{V_1, V_2, \ldots, V_n\}$
Output:	$\mathcal{F}_{video}$ and $\mathcal{F}_{audio}$
1.	# Data Preprocessing Phase
2.	for each video $V_i \in \mathbb{V}$:
3.	Split V_i into audio track A_i and image track I_i.
4.	$\mathcal{A}_i = D_{P\times M} \cdot \log\left(H_{M\times K} \cdot \left\lvert F_{K\times L}\left(w_L \odot f_{N,L}\left(P_\alpha\left(A_i\right)\right)\right)\right\rvert^2\right)$ # Process A_i using MFCC
5.	$\mathcal{D}_i = \sigma\left(W_d \cdot \sigma\left(W_e \cdot \mathcal{A}_i + b_e\right) + b_d\right)$
6.	$L_v = \max_{\forall I_i \in \mathbb{I}} L\left(I_i\right)$
7.	$\mathcal{V}_i = \left[I_i, Pad\left(L_v - L\left(I_i\right)\right)\right]$
8.	$F_i = \{f_{i,j} \mid f_{i,j} = \mathcal{V}_i\left(\frac{j}{N_v}\right), j \in [1, L_v \cdot N_v]\}$
9.	end for
10.	# Multimodal Anomaly Detection Phase
11.	for each pair $(\mathcal{D}_i, F_i)$:
12.	$\hat{y}_i^A = \mathcal{F}_{audio}\left(\mathcal{D}_i\right) = FC\left(Conv_k\left(\ldots Pool\left(Conv_1\left(\mathcal{D}_i\right)\right)\right)\right)$ # Audio anomaly detection
13.	$\hat{y}_i^I = \mathcal{F}_{video}\left(F_i\right) = FC\left(ConvLSTM_k\left(\ldots ConvLSTM_2\left(ConvLSTM_1\left(F_i\right)\right)\right)\right)$ # Video anomaly detection
14.	$\hat{y}_i^c = softmax\left(W_B \cdot \sigma\left(W_A \cdot \mathcal{F}_{audio}\left(\mathcal{D}_i\right) + W_I \cdot \mathcal{F}_{video}\left(F_i\right)\right) + b_B\right)$ # Integrated detection
15.	$\min_{\forall i} \mathcal{L}_{total} = \min_{\forall i}\left(\beta \cdot \mathcal{L}_{audio} + \gamma \cdot \mathcal{L}_{video} + \left(1-\beta-\gamma\right) \cdot \mathcal{L}_{combined}\right)$ # Minimization total loss
16.	end for
17.	return $\mathcal{F}_{video}$ and $\mathcal{F}_{audio}$.

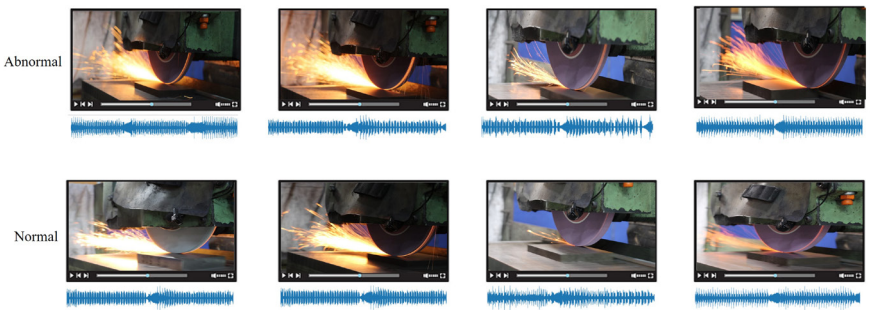


Figure 2. Sample videos from the Grinding Wheels dataset

5 Performance Evaluation

This section evaluates the proposed AMAD against single-modal methods, such as using CNN for audio and ConvLSTM for image recognition in detecting abnormal operations of grinding wheels. Relying solely on CNN for audio is vulnerable to environmental noise, while single-modal data fails to fully capture the characteristics of abnormal operations. Similarly, 3DCNN-based image recognition overlooks audio features and struggles with lighting variations and occlusions. The proposed AMAD addresses these issues by integrating both audio and video modalities, providing a comprehensive a comprehensive understanding of abnormalities and improving detection accuracy and robustness. This multimodal fusion approach outperforms single-modal methods, especially in complex environments.

5.1 Dataset

The evaluation utilizes a custom dataset known as the Grinding Wheels dataset, developed in collaboration with this study. Figure 2 shows the sample videos from the Grinding Wheels dataset. This dataset includes 328 video clips, with 175 normal and 153 abnormal operations. The dataset is split into 80% for training and 20% for testing.

5.2 Simulation Results

Figure 3 compares the training and validation losses across epochs, ranging from 0 to 50. Both the training and validation losses decrease and approach stable values as the number of epochs increases. This indicates that when the number of epochs increases, the model becomes more proficient at learning from the training data, leading to stabilized performance on the validation data. The initial rapid decrease in losses suggests efficient early learning, while the later stabilization reflects the model's convergence to an optimal state.

Figure 4 shows the training and validation accuracy across 50 epochs, demonstrating significant improvements as epochs increase, highlighting the effectiveness of AMAD in learning and parameter optimization.

Figure 5(a), Figure 5(b), and Figure 5(c) compare the proposed AMAD, video using 3DCNN, and audio using CNN in terms of precision, recall, and F1-Score across epochs (5–25) and dataset sizes (50–200). All models show a common trend that the precision, recall, and F1-Score increase with the number of epochs. The reason is that a higher number of epochs allows the models more opportunities to learn and refine their parameters, leading to better performance. Additionally, the precision, recall, and F1-score also improve as the size of the dataset increases. This is because a larger amount of training data enhances the models' learning capacities and enables them to more effectively capture the inherent features within the dataset. In comparison, the proposed AMAD outperforms the other methods due to its use of MFCC for precise audio preprocessing and multimodal fusion of audio and video data. This approach effectively detects anomalies in grinding wheel operations, achieving higher accuracy than single-modal methods.

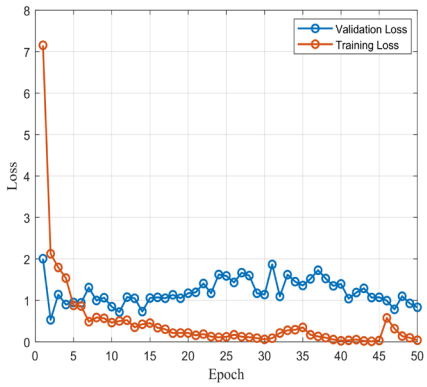


Figure 3. The loss of training and validation

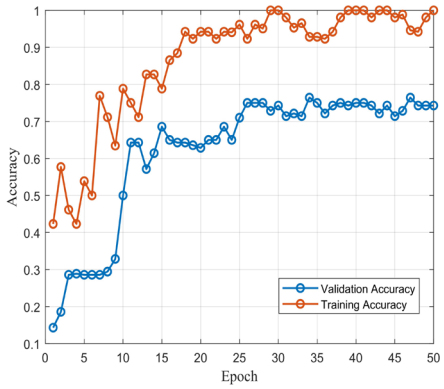
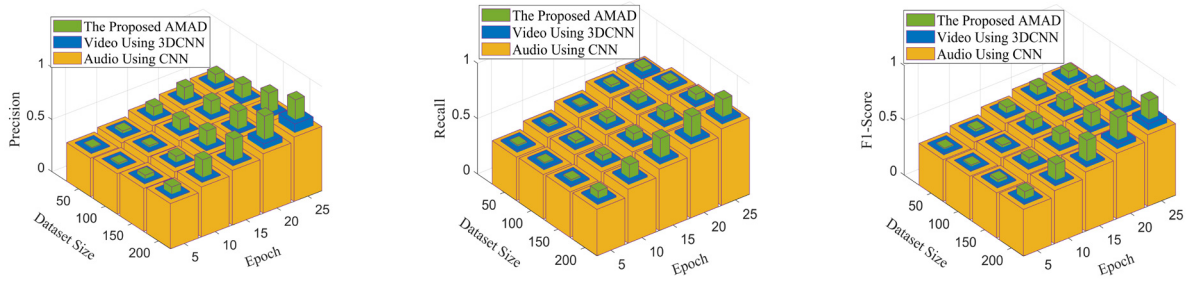


Figure 4. The accuracy of training and validation



(a) Precision (b) Recall (c) F1-Score

Figure 5. Comparison of different mechanisms by varying epoch and dataset size

Additionally, to emphasize the effectiveness of the proposed AMAD mechanism, Figure 6 employs the Friedman Test statistical analysis method to compare it with video using 3DCNN and audio using CNN mechanisms. The y-axis represents the F1-Score, while the x-axis lists these three mechanisms. The results are depicted as box plots for each mechanism, facilitating a comparison of their performance distributions. The Friedman Test yielded a chi-square value of 10.00 and a p-value of 0.00674. The p-value is below the conventional significance threshold of 0.05, indicating statistically significant differences among the three mechanisms. As shown in Figure 6, the proposed AMAD mechanism outperforms the other two mechanisms, demonstrating higher median and overall F1-Score, thus highlighting its superior performance in detecting anomalies.

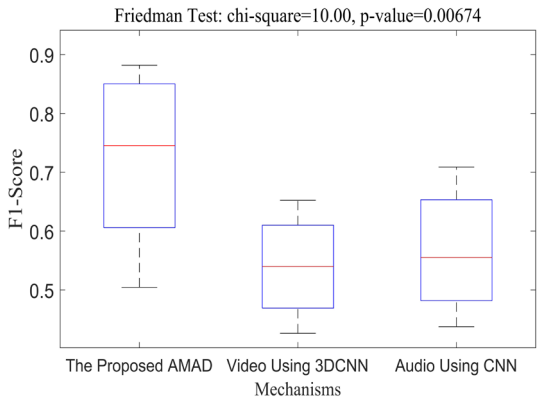


Figure 6. The results of the Friedman test using different mechanisms

Table 3 presents an ablation study on the Grinding Wheels dataset to evaluate the effectiveness of each module in AMAD, measured by F1-Score. The baseline, using only Audio with CNN, achieves an F1-Score of 0.561. Adding MFCC improves performance to 0.652, demonstrating efficient audio feature extraction. Incorporating Video with ConvLSTM raises the score to 0.729, and adding video preprocessing further increases it to 0.765. Combining audio and video modalities achieves the highest F1-Score of 0.836, highlighting the significant advantage of multimodal detection for improved anomaly detection accuracy.

Table 3. The ablation study of the proposed AMAD on the self-collected dataset

Modules	F1-Score
Audio using CNN	0.561
+ MFCC	0.652
+ Video using ConvLSTM	0.729
+ Video preprocessing	0.765
+ Combined multimodal detection	0.836

6 Conclusion

This paper presents AMAD, an anomaly detection mechanism for grinding wheel operations that leverages multimodal data to accurately identify abnormalities in operation videos. AMAD preprocesses audio using MFCC and AutoEncoder and video through segmentation techniques to ensure data readiness. It then analyzes audio with CNN and video with ConvLSTM, integrating both modalities to effectively classify operations as normal or abnormal. The results demonstrate the potential of the proposed AMAD to significantly improve the detection of abnormalities in industrial processes, offering a promising solution for enhancing operational efficiency and safety. Future work will focus on optimizing computational efficiency for faster processing and enabling real-time detection capabilities.

Acknowledgments

This work was supported in part by the Smart Home and Applied Industry Innovation Team, Higher Education Research Program Project under Grant Nos. 2022AH010067 and 2023AH051609, Anhui Outstanding Youth Fund Project under Grant No. 2022AH030109 and Anhui Provincial Natural Science Foundation under Grant No. 2408085MF177.

References

[1] F. L. Scudo, E. Ritacco, L. Caroprese, G. Manco, Audio-based anomaly detection on edge devices via self-supervision and spectral analysis, *Journal of Intelligent*

Information Systems, Vol. 61, No. 3, pp. 765-793, December, 2023.
[2] K. Wilkinghoff, F. Kurth, Why do angular margin losses work well for semi-supervised anomalous sound detection? *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp. 608-622, November, 2023.
[3] J. Wu, F. Yang, W. Hu, Unsupervised anomalous sound detection for industrial monitoring based on ArcFace classifier and gaussian mixture model, *Applied Acoustics*, Vol. 203, Article No. 109188, February, 2023.
[4] K. Jagadeeshwar, T. Sreenivasarao, P. Pulicherla, K. N. V. Satyanarayana, K. M. Lakshmi, P. M. Kumar, ASERNet: Automatic speech emotion recognition system using MFCC-based LPC approach with deep learning CNN, *International Journal of Modeling, Simulation, and Scientific Computing*, Vol. 14, No. 4, Article No. 2341029, August, 2023.
[5] F. Zou, X. Li, Y. Li, S. Sang, M. Jiang, H. Zhang, GOL-SFSTS based few-shot learning mechanical anomaly detection using multi-channel audio signal, *Knowledge-Based Systems*, Vol. 284, Article No. 111204, January, 2024.
[6] S. Kulkarni, H. Watanabe, F. Homma, Self-supervised audio encoder with contrastive pretraining for respiratory anomaly detection, *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, Rhodes Island, Greece, 2023, pp. 1-5.
[7] M. Wang, Q. Mei, X. Song, X. Liu, R. Kan, F. Yao, J. Xiong, H. Qiu, A machine anomalous sound detection method using the lMS spectrogram and ES-MobileNetV3 network, *Applied Sciences*, Vol. 13, No. 23, Article No. 12912, December, 2023.
[8] M. A. Hossan, S. Memon, M. A. Gregory, A novel approach for MFCC feature extraction, *4th International Conference on Signal Processing and Communication Systems*, Gold Coast, QLD, Australia, 2010, pp. 1-5.
[9] S. Sardari, B. Nakisa, M. N. Rastgoo, P. Eklund, Audio based depression detection using Convolutional Autoencoder, *Expert Systems with Applications*, Vol. 189, Article No. 116076, March, 2022.
[10] A. Ustubioglu, B. Ustubioglu, G. Ulutas, Mel spectrogram-based audio forgery detection using CNN, *Signal, Image and Video Processing*, Vol. 17, No. 5, pp. 2211-2219, July, 2023.
[11] H. Gao, S. Kuenzel, X. Zhang, A hybrid ConvLSTM-based anomaly detection approach for combating energy theft, *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, Article No. 2517110, August, 2022.
[12] A. Durairaj, E. S. Madhan, M. Rajkumar, S. Shameem, Optimizing anomaly detection in 3D MRI scans: The role of ConvLSTM in medical image analysis, *Applied Soft Computing*, Vol. 164, Article No. 111919, October, 2024.
[13] C. Lee, T. Yang, Z. Chen, Y. Su, Y. Yang, M. R. Lyu, Heterogeneous anomaly detection for software systems via semi-supervised cross-modal attention, *IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, Melbourne, Australia, 2023, pp. 1724-1736.
[14] P. Wang, X. Zhang, Z. Cao, Z. Chen, MADMM: microservice system anomaly detection via multi-modal data and multi-feature extraction, *Neural Computing and Applications*, Vol. 36, No. 25, pp. 15739-15757, September, 2024.
[15] H. Liu, X. Huang, M. Jia, T. Jia, J. Han, Z. Wu, Y. Li, UAC-AD: Unsupervised adversarial contrastive learning for anomaly detection on multi-modal data in microservice systems, *IEEE Transactions on Services Computing*, Vol. 17, No. 6, pp. 3887-3900, November-December, 2024.
[16] C. Feng, Z. Chen, A. Owens, Self-supervised video forensics by audio-visual anomaly detection, *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver*, BC, Canada, 2023, pp. 10491-10503.
[17] J. Gao, H. Yang, M. Gong, X. Li, Audio–visual representation learning for anomaly events detection in crowds, *Neurocomputing*, Vol. 582, Article No. 127489, May, 2024.
[18] R. Zhao, B. Du, L. Zhang, Hyperspectral anomaly detection via a sparsity score estimation framework, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55, No. 6, pp. 3208-3222, June, 2017.
[19] H. Su, Z. Wu, A. Zhu, Q. Du, Low rank and collaborative representation for hyperspectral anomaly detection via robust dictionary construction, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 169, pp. 195-211, November, 2020.

Biographies

Qiaoyun Zhang received her Ph.D degree in Computer Science and Information Engineering from Tamkang University. She is currently a Lecturer with the School of Artificial Intelligence, Chuzhou University, Chuzhou, Anhui, China. Her current research interests focus on artificial intelligence.

Hsiang-Chuan Chang is currently an Assistant Professor in the Deparment of Transportation Management, Tamkang University, New Taipei City, Taiwan. His current research interests include light rail transit system and artificial intelligence.

Chia-Ling Ho received her Ph.D. in Management Sciences from Tamkang University and currently holds the position of associate professor at National Taipei University of Nursing and Health Sciences, Taipei, Taiwan in 2008. Her research is dedicated to information systems development, analysis and the application of management science methodologies.

Huan-Chao Keh is currently a full professor in the Department of Computer Science and Information Engineering at Tamkang University, Taiwan. He has been the President of Tamkang University since August 1, 2018. His current research interests include Data Mining, Internet of Things, Artificial Intelligence and Clinical Medical Information Systems.

Diptendu Sinha Roy received the Ph.D. Eng. degree from Birla Institute of Technology, Mesra, India in 2010. He is currently a professor with the Department of Computer Science Engineering, National Institute of Technology Meghalaya, India. His current research interests include 5G, software reliability, and machine learning, big data.