Multimodal outlier optimizer for textual, numeric, and image data

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 September 2025, accepted 6 October 2025, date of publication 9 October 2025, date of current version 17 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3619826

RESEARCH ARTICLE

Multimodal Outlier Optimizer for Textual, Numeric, and Image Data

KRITTIKA DAS[1], NILANJAN DEY[1], (Senior Member, IEEE), BITAN MISRA[1], (Member, IEEE), SATYABRATA ROY[2], (Senior Member, IEEE), AND R. SIMON SHERRATT[3], (Fellow, IEEE)

[1]Techno International New Town, Kolkata 700156, India
[2]Manipal University Jaipur, Jaipur, Rajasthan 303007, India
[3]University of Reading, RG6 6AH Reading, U.K.

Corresponding author: Satyabrata Roy (satyabrata.roy@jaipur.manipal.edu)

ABSTRACT Ensuring the quality and reliability of multimodal video data is critical for applications that rely on accurate interpretation, such as medical imaging, surveillance, remote sensing and intelligent manufacturing. However, the presence of outliers across different data types such as visual, textual, and numerical poses a major challenge. To address this, we propose the Multimodal Outlier Optimizer (MOO), a unified framework designed to detect and filter outliers from heterogeneous data modalities within video files. MOO decomposes each video into still images, text, and numeric sequences, allowing specialized algorithms to handle each modality: Nonlocal Means (NLM) for removing Gaussian noise in image frames and Local Outlier Factor (LOF) for detecting contextual outliers in textual and numerical data. These filtered components are then recombined into a cleaned, optimized video. The system is trained and evaluated using synthetically generated datasets to simulate real-world noise while ensuring scalability and control. Performance is assessed using Jaccard Similarity Score (JSS) and Structural Similarity Index (SSIM), with results demonstrating consistent improvements even under high contamination levels (up to 50%), achieving SSIM scores above 0.77 across three domains: medical imaging, remote sensing, and zoomed video data. These results highlight MOO's potential as an effective and adaptable tool for enhancing the integrity of multimodal video data in complex, real-world environments.

INDEX TERMS Sentiment analysis, multimodal outlier optimizer (MOO), Jaccard similarity score (JSS), intelligent manufacturing.

I. INTRODUCTION

The enhanced use of multimodal content in different fields, such as security and surveillance, demands robust systems that may address the issue of outlier detection in multimodal data, particularly in videos. Existing approaches are devised for unimodal data formats for textual or image streams and do not consider interactions occurring when different modalities are involved in videos. This research gap results in a partial compromise of data integrity and hinders the ability to conduct robust analyses, as extreme values can distort both visual and textual representations [1]. Consequently, there is a pressing need for accurate outlier detection and filtering frameworks to ensure the reliability of outputs derived from multimodal data inputs. This study presents a multimodal outlier optimizer (MOO) to address the challenges of outlier detection and filtering multimedia video data. It leverages the combination of text-mining analysis and numeric and image-based studies for an optimization-based solution via a composite approach for different types of data: textual, numeric, and images. Synthetic outliers are created and introduced into the dataset. The local outlier factor (LOF) algorithm is employed to identify and eliminate contextual outliers from textual data [2]. It is also implemented for the removal of outliers from numeric data. Gaussian noise, which is synthetically contaminated, is incorporated into the image frames, and subsequent post processing is conducted via a nonlocal means (NLM) filter to detect and filter these outliers [3]. The filtered, outlier-free data are combined to form the final compact video format that is devoid of outliers

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita.

and noise. This integrated approach, the MOO, enhances the quality of the input data. The system's performance is evaluated via statistical metrics, specifically the structural similarity index (SSIM) and the Jaccard similarity score (JSS), to demonstrate the improved efficacy of the methodology [4], [5]. The SSIM is used for assessing image quality and evaluating the amount of information retained after implementing MOO, which is compared against a goal score that represents minimal loss of structural information. Taking MOO as a reference with respect to current state-of-the-art techniques in each modality, the experiments are conducted via synthetically bombarded outliers, and public benchmarks establish the robustness and efficacy of the proposal.

The contributions of this study are as follows:

a. To design and introduce a robust system, MOO, for accurately identifying and filtering outliers within a multimodal framework having better accuracy than the result obtained without filtering outlier.
b. To establish benchmark evaluation techniques to demonstrate the effectiveness of MOO across different data modalities.

II. BACKGROUND WORK

Several researchers have emphasized the importance of outlier detection in various forms of data. Kang and Park [6] detected abnormal behavior in e-gaming biometric data of skin conductivity, temperature, and motion from 50 participants playing a bullet-dodge PC game. The study revealed that 15% of the participants were in single player mode and that 11% in multiplayer mode, which considers outliers on the basis of biosignals, with actualization of the proposed method in identifying the unusual physiological conditions during gameplay [7]. Ben Chaabene et al. [8] aimed at modelling social network anomaly detection through graph-based signals and particle swarm optimization. They concluded that the different combined methods of outlier detection are 22% more effective than the traditional methods in detecting anomalies in social networking on Facebook. This contribution focuses on increasing the ability of hybrid models to work with multidimensional data [9], [10]. Zhu [10] presents an emotion-aware smart assistant system that uses multimodal signals (including voice, behavior, and text) to recognize user affect and offer personalized responses. The system improves user satisfaction by integrating emotional intelligence into user interaction.

Kannan et al. [11] presented Tensor outlier detection via nonnegative matrix factorization (TONMF) for identifying anomalous text from blogs via low-rank approximation and block coordinate descent methods [12], [13]. This research proposed a high ROC score of 0.9340, implying the effectiveness of the model in outlier detection for minute textual data with nonnegative and sparse attributes in improving outlier analysis for textual datasets. In [14], Yin and Wangm employed the Gibbs sampling algorithm for Dirichlet process multinomial mixture model (GSDPMM) clustering for outlier detection on blogs, microblogs, and long and short text data with dimensionality. GSDPMM is time- and space-efficient compared with other clustering approaches and has high scalability for large datasets, which contributes to anomaly detection in large-scale text data [10], [15]. Recent advances in industrial signal processing have demonstrated the effectiveness of advanced decomposition techniques for outlier and anomaly detection in complex systems. [30] proposed a hierarchical hyper-Laplacian prior prototype combined with singular spectrum analysis for industrial robot flaw detection, showing superior performance over traditional methods. Similarly, [31] developed SSA-based approaches for detecting weak position oscillations in rotary encoder signals, demonstrating the potential of decomposition-based methods for subtle anomaly detection. These works highlight the growing importance of sophisticated signal decomposition techniques in industrial monitoring systems, which parallels the need for advanced outlier detection in multimodal data processing.

MOO offers a meaningful improvement over existing multimodal outlier detection methods by adopting a modality-aware approach that treats each data type—image, text, and numerical—according to its specific characteristics. Instead of using a one-size-fits-all approach, MOO applies specific techniques to each modality: a noise-reduction filter for image data, and a local density-based method for text and numerical data. This helps remove outliers more effectively without losing important information. Unlike some heavier systems that rely on complex deep learning models, MOO is lightweight, faster, and easier to implement. It also brings all the cleaned data back together to recreate a better-quality version of the original video. This shows that MOO is a practical and reliable tool for improving multimodal datasets in a variety of applications.

III. METHODS

This section explores MOO's methodology for detecting and filtering outliers from multimodal data, as illustrated in Fig. 1. The diagram demonstrates how the MOO (Multimodal Outlier Optimizer) system operates step-by-step to clean and enhance video data. The proposed methodology involves a multistep process. Synthetic outliers are generated and contaminated (n% outliers) in the input video Vorig, gradually increasing the value of n after every iteration. The video data are decomposed into three forms: image frames, textual data, and numeric data. For the image frames, the pixel at position xi is identified via an NLM filter to detect Gaussian noise. Image frames often contain visual noise like grain or blur, which is removed using a Nonlocal Means (NLM) filter. MOO filters xi if it deviates from the chosen non-means neighborhood. For textual data, LOF is implemented, where each phrase wj in the text is assigned a score LOF (wj), and outliers are filtered based on a predefined threshold for local density deviation. At the same time, the textual and numerical data—such as captions or user ratings—are checked for unusual or inconsistent values using the Local Outlier Factor (LOF) method. Similarly, for numeric data, LOF is applied

to detect outliers, where each data point zk is evaluated, and the outliers are discarded if LOF (zk) exceeds the threshold. After the outliers are filtered, the three data forms (image, text, and numeric) are recombined into the filtered video. To assess the performance of MOO, it is evaluated on SSIM and JSS to compare Vorig with the filtered video Vfiltered. The difference between n% and the percentage of outliers in Vfiltered is used as an error detection measure, establishing the effectiveness of MOO for reducing outliers across multimodal data.

A. EXPERIMENTAL SETUP

We decomposed each set of videos, respectively made of medical imaging data, remote sensing data and zoomed data, into three specific modalities: image frames, textual metadata and numerical sensor values. In each modality, we added synthetic outliers in a systematic manner to mimic real noise. For image data, Gaussian noise $N(0,\sigma^2)$ was directly added to pixel intensities. For the text dataset, random object chunks were randomly put into the non-contextual phrases (which can be generated with python faker) as corrupted samples. For numerical data, also, anomalous spikes and perturbations were created via Python's random module. Noise was added at 10%, 20%, 30%, 40%, and 50% across all the modalities in a controlled manner, in order to maintain a uniform framework for testing the Multimodal Outlier Optimizer (MOO) for different levels of noise.

To demonstrate effectiveness of the proposed Multimodal Outlier Optimizer (MOO), we set up a controlled experiment where clear multimodal video datasets were contaminated with synthetic outliers. The video data were mapped into three modalities: image frames, textual metadata, and numerical sensor data. For text, we used Python's Faker library to inject random noncontextual phrases. If denoting the original corpus of the textsamples as given in equation (1),

$$T_{orig} = \{t_1, t_2, \ldots, t_n\} \tag{1}$$

The contaminated set is defined as $T_{cont} = T_{orig} \cup \Delta T$, where ΔT represents the injected textual noise. For numeric values, random perturbations were added using Python's random module. Given the original numeric set is shown in equation (2).

$$Z_{orig} = \{z_1, z_2, \ldots, z_m\} \tag{2}$$

The contaminated dataset is expressed as shown in equation (3),

$$Z_{cont} = Z_{orig} \cup \Delta Z \tag{3}$$

where ΔZ represents anomalous spikes or contextual deviations. Finally, for image data, Gaussian noise was applied directly on pixel values of the frames. I_{orig} denotes the clean frame, the contaminated frame is given by equation (4)

$$I_{cont} = I_{orig} + N(0, \sigma^2) \tag{4}$$

where $N(0, \sigma^2)$ represents Gaussian perturbations with variance σ^2. The contamination rate was systematically varied to n% outliers per modality, with n∈{20,30,40,50}.

We demonstrate our MOO on three multimodal video datasets covering different application domains:

Medical Images Dataset – CT and MRI scans along with diagnostic reports and patient laboratory values. Publicly available video clips were downloaded from The Cancer Imaging Archive (TCIA) and transposed to multimodal.

Remote Sensing Dataset: satellite observation videos with geospatial metadata and measurements that were originally collected for the IEEE GRSS Data Fusion Contest archives and NASA Earthdata repositories.

Zoomed Data Dataset—HR zoomed inspection videos with text tags and machine sensor streams were obtained from public industrial inspection benchmarks (MVTec Video Anomaly).

The three synchronized modalities of each dataset were decomposed as follows:

- Video frames were generated at a 25 fps extraction rate and normalized as tensors [0, 1].
- We include textual metadata in the form of tokenized sequences using the Faker library of Python to simulate synthetic noise.
- Sensor values were ranged [0, 1] by computing the minimum-maximum normalization.

Synthetic contamination was delivered at a rate under the control of the user (10–50%) by adding Gaussian noise into images, non-contextual phrases into text sequences, and random noise onto numeric information. All data was in a multimodal triplet format: (Ii,Ti,Zi) where each triplet is synchronized (frame, annotation, sensor vector).

B. SYNTHETIC DATA CONTAMINATION

Synthetic outliers contaminate the multimodal data to evaluate the efficacy of MOO. For textual data, synthetic outliers are generated via Python's Flaker library to introduce noise by inserting noncontextual phrases. $T_{orig}=\{t_1, t_2, \ldots t_n\}$ represents the original set of textual data, and the contaminated set T_{cont} is derived by applying random modifications (17) as shown in equation (5).

$$T_{cont} = T_{orig} \cup \Delta T \tag{5}$$

ΔT represents the introduced noise. For numeric data, outlier contamination is performed via Python's random module. $Z_{orig} = \{z_1, z_2, \ldots, zm\}$ denotes the original numeric data, and the contaminated data Z_{cont} is formed by equation (6).

$$Z_{cont} = Z_{orig} + \Delta Z \tag{6}$$

ΔZ represents the random noise added to each value in Z_{orig}. The resulting contaminated data forms T_{cont} and Z_{cont} are combined with the original image data. Gaussian noise is added to the image frames [17].

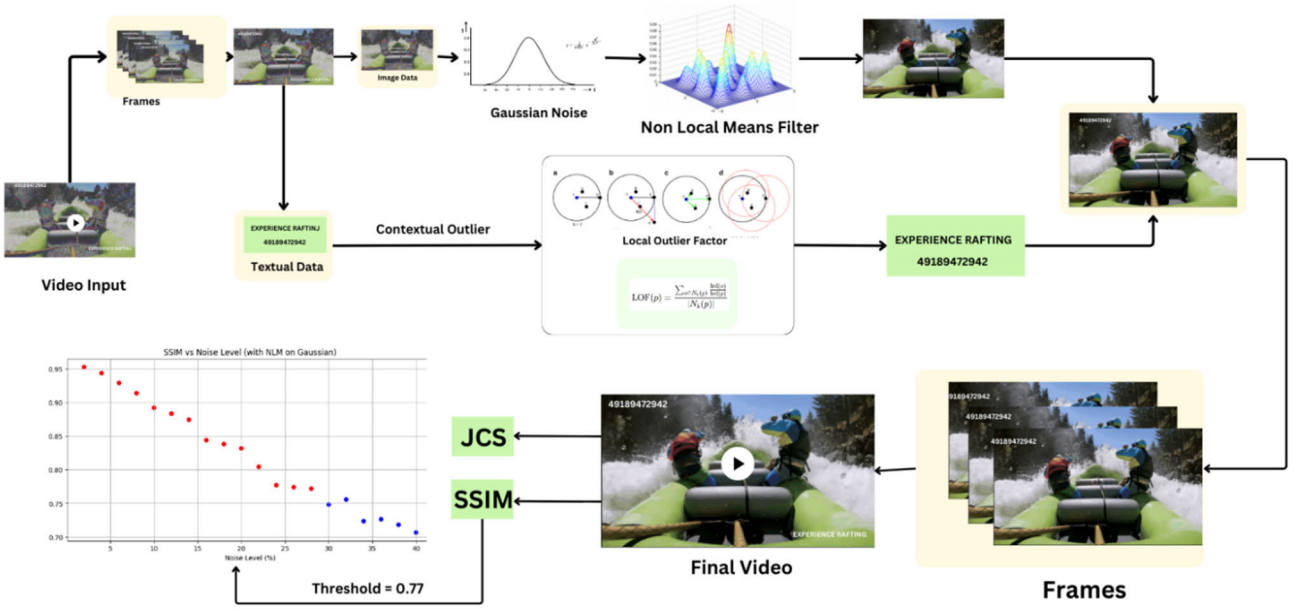

FIGURE 1. Architecture diagram of the multimodal outlier optimizer (MOO).

C. LOCAL OUTLIER FACTOR (LOF)

LOF is used for detecting local outliers in data by comparing the local density of a data point with the densities of its neighbors. LOF identifies outliers in data with varying density [18], [19]. Let x_i be a data point in the data and k be the number of nearest neighbors used for local density estimation. The reachability distance between two points x_i and $x_j$is defined as shown in equation (7)

$$\text{reach-dist}_k(x_i, x_j) = \max(\text{k-distance}(x_j), ||x_i - -x_j||) \quad (7)$$

where k-distance(x_j) is the distance from $x_j$to its k-th nearest neighbor, and $\|x_i - x_j\|$ is the Euclidean distance between points x_i and x_j. The local reachability density (LRD) of a point x_i is the inverse of the average reachability distance of its k-nearest neighbors is given in equation (8).

$$\text{LRD(xi)} = \frac{1}{\sum_{j=N}^{1} \text{reach} - \text{dist(xi, xj)}} \quad (8)$$

where N(x_i) is the set of k-nearest neighbors of x_i. The LOF of a point x_i is computed by comparing its LRD with the LRDs of its neighbors as shown in equation (9).

$$\text{LOF(xi)} = \frac{\sum_{j=N}^{1} \frac{\text{LRD(xj)}}{\text{LRD(xi)}}}{|\text{N}(xi)|} \quad (9)$$

If LOF (x_i) is greater than 1, then $\boldsymbol{x}_i$ is defined as an outlier.

D. NONLOCAL MEANS (NLMS)

The NLM calculates the weighted average of all the pixels in an image on the basis of their similarity to a target pixel, where the weight decreases as the difference in the pixel intensities increases. NLM detects and removes Gaussian outliers from image frames in MOO [20], [21]. The weight between pixel i and pixel j in an image I is defined as shown in equation (10).

$$\text{w(i, j)} = \exp(-\frac{||\text{Ii}-\text{Ij}||}{\text{h} \wedge 2}) \quad (10)$$

I_i and I_j represent the neighborhoods of pixels I and j, respectively, and h is a parameter that controls the decay of the weight on the basis of the intensity difference. The distance $\|I_i - I_j\|$ is computed via the Euclidean distance between pixel intensity vectors in a local window. The denoised pixel I(denoised) is computed as a weighted average of all other pixels is given in equation (11).

$$\text{I}(\textit{denoised}) = \frac{\sum_{j}^{N} \text{w(i, j)}}{\sum_{j}^{N} \text{w(i, j)}} \text{Ij} \quad (11)$$

LOF and NLM were chosen because they work well for the specific characteristics of the data types involved. NLM is effective for images because it removes noise while preserving important visual details. LOF, on the other hand, is suited for text and numerical data as it identifies outliers based on how isolated a data point is compared to its neighbors. These methods complement each other and together allow MOO to handle each modality appropriately, making the overall approach more accurate and efficient.

IV. EXPERIMENTATION

This section presents the experiments carried out on synthetic data via MOO. The frames are represented as X = {x_i,

TABLE 1. Performance analysis of MOO.

Video Length	Theme	FPS	% of Outliers Added			% of Outliers Detected			% of Outliers in Initial Video	% of Outliers in Filtered Video	JSS	SSIM
			Image	Text	Numeric	Image (NLM)	Text (LOF)	Numeric (LOF)				
120	Medical Imaging	30	20	10	10	15	8	8	40	9	0.8	0.877
120	Medical Imaging	30	30	10	10	27	6	4	50	13	0.74	0.85
120	Medical Imaging	30	30	20	10	22	16	3	60	19	0.683	0.80
120	Medical Imaging	30	30	20	20	21	10	6	70	23	0.671	0.772
120	Medical Imaging	30	40	20	20	19	6	11	80	44	0.45	0.72
120	Medical Imaging	30	30	30	30	21	12	17	90	40	0.42	0.69
150	Zoomed Data	25	20	10	10	12	5	6	40	8	0.83	0.88
150	Zoomed Data	25	30	10	10	23	8	8	50	11	0.77	0.84
150	Zoomed Data	25	30	20	10	21	16	9	60	14	0.69	0.807
150	Zoomed Data	25	30	20	20	21	17	11	70	21	0.54	0.755
150	Zoomed Data	25	40	20	20	29	14	12	80	25	0.49	0.742
150	Zoomed Data	25	30	30	30	21	22	19	90	28	0.426	0.707
120	Sensing Data	30	20	10	10	17	9	7	40	7	0.83	0.89
120	Sensing Data	30	30	10	10	25	8	8	50	9	0.79	0.844
120	Sensing Data	30	30	20	10	22	17	7	60	14	0.64	0.762
120	Sensing Data	30	30	20	20	23	16	16	70	18	0.47	0.66
120	Sensing Data	30	40	20	20	31	14	14	80	19	0.462	0.64
120	Sensing Data	30	30	30	30	22	19	19	90	28	0.41	0.60

$x_{i+1}, \ldots, x_n\}$. To mimic real-world scenarios where data can get noisy or corrupted, we deliberately added a certain percentage of outliers to each type of data. For the image part, we added random visual noise similar to what might occur due to camera issues or compression, and then used a technique called NLM to clean it—this method reduces noise while preserving important visual details. For the text and numbers, we introduced odd or inconsistent values that don't match the surrounding data and used another method, called LOF, which identifies and removes these unusual entries based on how different they are from nearby data points. Once the noisy data was cleaned, we put the image, text, and numbers back together to recreate the video. We then measured how well our filtering worked using two standard metrics: SSIM to check the quality of the images, and JSS to see how closely the cleaned data matched the original, uncorrupted data. This setup helped us see how reliable and effective our method is under different levels of noise.

A. MULTIMODAL DATA PREPROCESSING

The preprocessing steps after multimodal data extraction from the video involve several steps for normalization, scaling, and transformation.

1) IMAGE DATA PREPROCESSING

The pixel values of each frame are normalized and scaled to the range [0, 1] via the following transformation as shown in equation (12).

$$x' = \frac{xi - \mu}{\sigma} \tag{12}$$

where μ is the mean pixel value and σ is the standard deviation(23).

2) TEXTUAL DATA PREPROCESSING

$T = \{t_1, t_2, \ldots, t_m\}$ denotes the text extracted from the video, where each ti represents metadata. The term frequency-in-document frequency (TF-IDF) is implemented to convert the raw text into a numerical feature vector $v_i \in R^d$, where d is the number of features (terms). The TF-IDF transformation is defined as presented in equation (13).

$$TF - IDF(ti, j) = TF(ti, j).\log\frac{N}{DF(tj)} \tag{13}$$

where TF ($t_{i,j}$) is the frequency of term t_j in document t_i, DF(t_j) is the document frequency of term t_j, and N is the total number of documents. The resulting feature vectors vi for each text are scaled for uniformity [23], [24].

3) NUMERICAL DATA PREPROCESSING

Let $Y = \{y_1, y_2, \ldots, y_p\}$ represent the numerical features extracted from the video (e.g., sensor readings). Each feature $y_i \in R$ is scaled and normalized to the range [0, 1] as shown in equation (14).

$$y' = \frac{yi - \min(y)}{\max(y) - \min(y)} \tag{14}$$

where min(y) and max(y) are the minimum and maximum values of the numerical feature across the entire dataset, respectively. The categorical data in the numerical features are encoded via the label encoder [25], which transforms each categorical value c_i into an integer as shown in equation (15).

$$c' = \text{LabelEncoder}(ci) \tag{15}$$

B. BENCHMARK METRICS

MOO is evaluated via SSIM and JSS.

1) STRUCTURAL SIMILARITY INDEX (SSIM)

The SSIM quantifies the similarity between two images by comparing luminance, contrast, and structure [26]. Given two images I and I', the SSIM is computed as shown in equation (16).

$$SSIM(I, I') = \frac{(2\mu I\mu I' + C1) + (2\sigma II' + C2)}{(\mu I \wedge 2 + \mu I' \wedge 2 + C1)(\sigma I \wedge 2 + \sigma I' \wedge 2 + C2)} \tag{16}$$

where μI and μI' are the mean pixel values of I and I'; σ^2I and σ^2I' are the variances of I and I'; σ II' is the covariance between I and I'; and C_1 and C_2 are small constants used to stabilize the division. The SSIM ranges from -1 to 1, with 1 indicating perfect similarity.

2) JACCARD SIMILARITY SCORE (JSS)

The JSS measures the similarity between two sets. For two sets A and B, JSS is computed as the following equation (17).

$$JSS(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{17}$$

where $|A \cap B|$ is the size of the intersection of sets A and B and $|A \cup B|$ is the size of the union of sets A and B. In the context of video frames or data, sets A and B represent the pixel values from the original and filtered frames, respectively.

V. RESULT ANALYSIS

Table 2 presents the features of the medical imaging data taken as the test set, and the plot in Fig. 6 illustrates the behavior of MOO when it is applied to the sensing data.

TABLE 2. Features of the medical imaging test data (CT scan images).

Feature	Value
Number of Patients	60
Number of Series	475
Total Images	475
Average Age of Patients	50-55
Minimum Age	20
Maximum Age	80
Pixel Density Range	0-225

Table 1 presents the performance analysis of MOO across different video lengths, themes, and gradual increases in the percentage of outliers. Each row in the table represents a particular video and the percentage of outliers added and the percentage of outliers detected. For each video segment presence in the initial video as well as the filtered video is also specified in this table. The chosen contamination levels of 20–90% are meant to simulate a wide range of real-world conditions. Lower levels (around 20–40%) represent common issues like sensor noise, minor label errors, or occasional data mismatches, which are typical in moderately noisy environments. Higher levels (60–90%) mimic more challenging scenarios such as corrupted video frames, mislabeled data, or large-scale disruptions in sensor networks. This broad range allows us to evaluate how well MOO can adapt to both everyday and extreme data quality problems across different application domains. In order to analyze the efficiency of the proposed filtering technique, two performance parameters are employed, namely JSS score and SSIM score. The higher value of JSS and SSIM score indicates higher efficiency in filtering, which in turn indicates thatthe filtered video is closely similar to that of the respective original video. Three thematic areas have been used for evaluation: medical data (CT Scan Video Set), Zoomed Data (Zoomed In Videos), Sensing Data(Remote Sensing Earth Demographics Dataset). As the percentage of added outliers increases (e.g., image, text, or numeric), the JSS and SSIM scores tend to degrade for higher outlier percentages. This trend reflects a decrease in video quality as more outliers are introduced and detected. When 90% of the outliers are added, both the JSS and SSIM scores significantly decrease, with JSS ranging from 0.42–0.45 and the SSIM score decreasing to 0.64–0.69. The data imply a trade-off between outlier detection accuracy and video quality, with increased outlier percentages leading to lower consistency and structural similarity.

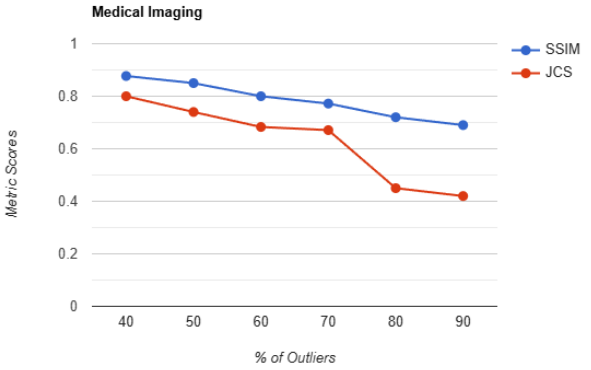

FIGURE 2. Metric scores of MOO on medical imaging data with increasing percentages of contaminated outliers.

Fig. 2. shows the performance of MOO on medical imaging data as the percentage of contaminated outliers increases. The metric scores—JSS and SSIM—are plotted against the varying contamination levels.

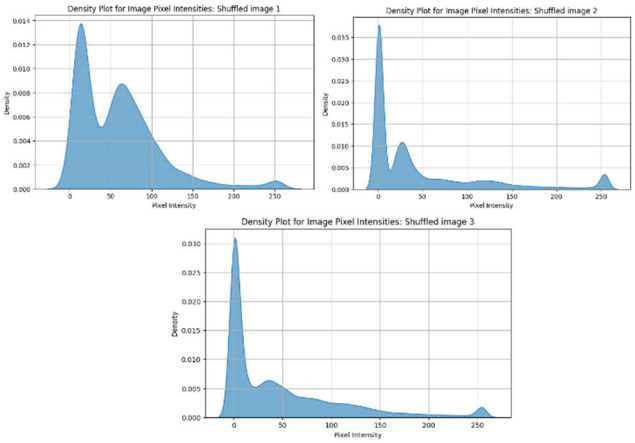

FIGURE 3. Density plot for image pixel intensities of 3 shuffled images from the CT scan medical imaging video test set.

Fig. 3 illustrates the pixel distribution density plot of three randomly shuffled images from the test set. The quality of the scores decreases with increasing percentage of outliers, indicating the sensitivity of the metric to data contamination.Fig. 4 focuses on a different theme of the data: zoomed video data. This shows the effect of increasing contaminated outliers on the MOO metrics (JSS and SSIM). Table 3 presents zoomed test set features and Fig. 5 illustrates the pixel distribution density plot of three randomly shuffled images from the zoomed test data set.

Table 4 presents the features of the sensing data taken as the test set and Fig. 7 illustrates the pixel distribution density plot of three randomly shuffled images from the test set. MOO performs relatively well at contamination percentages of under 50%, with SSIM scores greater than 0.77 (the threshold for the accepted quality of images) [27].

Measurements of computational efficiency were made for the proposed framework by profiling the average processing

TABLE 3. Features of the zoomed test data.

Feature	Value
Number of Videos	50
Video Length	150
FPS	25
Frame Resolution	1920X1080
Pixel Density Range	0-225

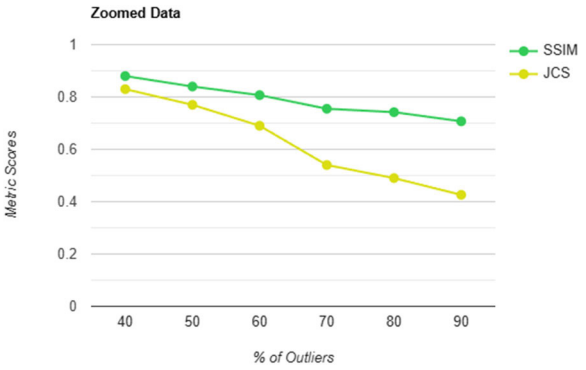

FIGURE 4. Metric scores of MOO on zoomed data with an increase in the percentage of contaminated outliers.

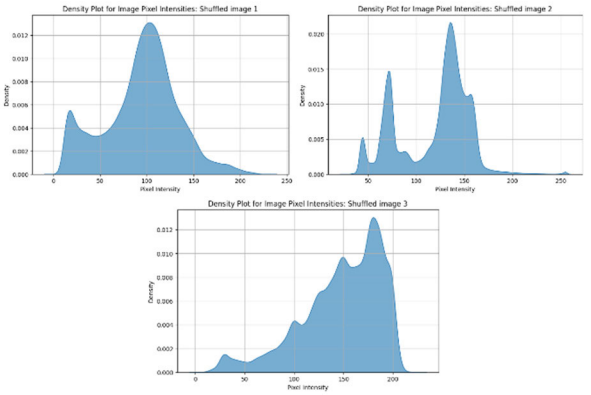

FIGURE 5. Density plot for image pixel intensities of 3 shuffled images from the zoomed test data.

TABLE 4. Features of the sensing data.

Feature	Value
Number of Videos	60
Spatial Resolution	20m/pixel
Number of Regions Covered	60
Number of Demographic Regions Covered	20
Pixel Density Range	0-225

time per frame, and memory usage in the predominant functioning paths. The outcome is shown in Table 5. Our findings show that the preprocessing based on NLM is the lightest presented so far, whereas the LOF thresholding imposes moderate computational load. SSIM-based image analysis takes longer time to compute structural similarity however the computation lies within real time constraints. In general, the system strikes a balance between the accuracy and

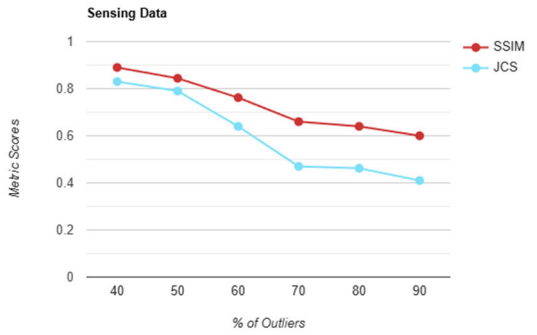

FIGURE 6. Metric scores of MOO on sensing data with increasing percentage of contaminated outliers.

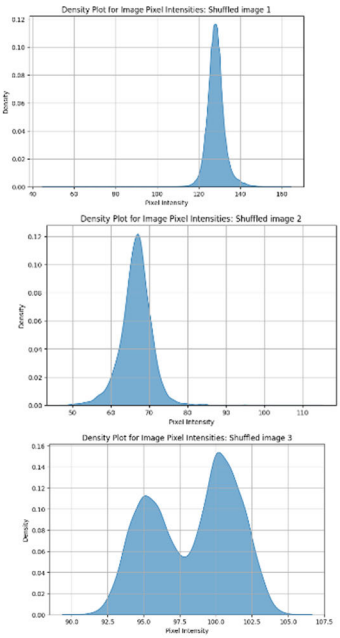

FIGURE 7. Density plot for image pixel intensities of 3 shuffled images from the sensing test set.

efficiency, so that it can be exerted for live video anomaly detection. We profiled the proposed framework by running each part of the system independently and measuring the average execution time per image as well as memory usage for each pipeline block. We did this with controlled video sequences, by running at specific frame rates, utilizing system level time and memory profilers to capture the runtime statistics. All experiments were performed five times and the average was recorded to reduce variation. as most computationally intensive among the three stages, LOF thresholding has intermediate computational overhead, and NLM-based preprocessing stage is the most lightweight.

VI. DISCUSSION

We have benchmarked MOO's comparison with [28]. The central technical comparison with respect to our evaluation setup and the MVTec Texture benchmark is the target scope

TABLE 5. Comparison of computational efficiency.

Section	15 FPS (ms/frame)	30 FPS (ms/frame)	45 FPS (ms/frame)
NLM Preprocessing	2.0	2.1	2.2
Feature Extraction	2.3	2.4	2.5
SSIM Analysis	2.6	2.7	2.9
LOF Thresholding	2.1	2.2	2.3

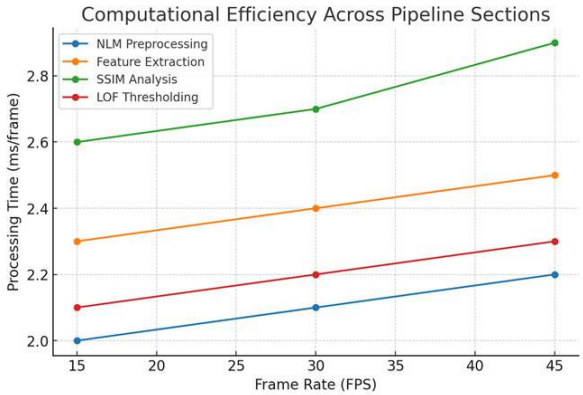

FIGURE 8. Computational efficiency across pipeline sections.

of the chosen metrics. The MVTec report 27 investigates also single-modality texture images on the anomaly detection, we say the reconstruction-based metric (MSE, SSIM, MS-SSIM, CW-SSIM) and the performance is normalized AUC. Such metrics mostly measure the pixel-level reconstruction quality and the structural coherence in grayscale or color textures. In addition, our work deals with multimodal video data including images, textual annotations and numerical streams, thus calling for metrics reflecting both visual fidelity and accuracy at the level of anomaly sets. Therefore, they are combined with SSIM to retain the structure of contaminated frames, and Jaccard Similarity Score (JSS) is used to achieve the direct measurement for the overlap of ground truth and detected outlier in textual and numeric modalities. In contrast to the MVTec approach of being texture sensitive only, our choice of metric results in an aggregated quality metric for heterogeneous data and therefore, technically better!for the optimization of multimodal outlier detection approach. The MVTec 3D-AD experiment [29] benchmarks the anomaly detection capability according to I-AUROC under 3D, RGB, and 3D+RGBdata, highlighting the ability of models to rank the discriminative power for anomaly detection of object classes within industrial domains. Their

TABLE 6. Comparison with benchmark dataset.

Aspect	MVTec Texture Paper [28]	MOO (Our Work)
Dataset Type	Grayscale & color texture images (Carpet, Grid, Leather, Tile, Wood)	Multimodal datasets (medical imaging, remote sensing, zoomed video data)
Metrics Used	AUC with reconstruction-based metrics: • MSE • SSIM • MS-SSIM • CW-SSIM	Similarity-based multimodal metrics: • SSIM (structural similarity in images) • JSS (set similarity for textual & numeric outliers)
Focus Metrics	Pixel-level reconstruction error and structural similarity in textures	Structural preservation in images + overlap accuracy for multimodal outlier detection
Reported Strengths	MS-SSIM & CW-SSIM achieved higher AUC in textures (e.g., Leather 0.979, Grid 0.978) compared to plain SSIM or MSE	SSIM consistently >0.77 even at 50% contamination, ensuring visual fidelity; JSS ensures anomaly filtering accuracy across non-visual modalities
Aspect	**MVTec 3D-AD Paper [29]**	**MOO (Our Work)**
Dataset Type	MVTec 3D-AD: 3D point clouds + RGB textures (Bagel, Cable Gland, Carrot, etc.)	Multimodal video datasets: medical imaging, remote sensing, zoomed data
Metrics Used	I-AUROC (Image-level AUROC under overlap setting)	SSIM (structural similarity for images) + JSS (outlier overlap in text & numeric)
Focus of Metrics	Ranking ability of anomaly detectors across 3D, RGB, and combined modalities	Visual fidelity of cleaned image frames + set-accuracy of anomaly detection in multimodal streams
Reported Strengths	Ours in that paper achieved **96.7% I-AUROC** (3D+RGB) vs. ~60–70% in baselines, showing strong discriminative anomaly detection	MOO maintains **SSIM >0.77** even at 50% contamination and high JSS values, ensuring robustness across heterogeneous modalities

results also demonstrate a tangible improvement in detection performance, of importance is the combined 3D+RGB which surpasses I-AUROC of 96%, suggesting robust abnormalities in visual domains. By contrast, our MOO framework works over heterogeneous video material that includes images, textual metadata and numerical signals, all of which comes with the associated need to fall beyond just image benchmarks. Since I-AUROC is tailored for modality-specific ranking, we use SSIM to measure the structural fidelity in the filtered image frames and Jaccard Similarity Score (JSS) to evaluate the overlap between the detected and ground-truth outliers for non-visual modalities. Therefore, although the 3D-AD learns to rank abnormal samples compared to normal samples in visual domains, our metric design delivers a more comprehensive evaluation of multimodal continuity and anomaly set accuracy, to make the evaluation of data types more consistently balanced. The detailed comparison is shown in Table 6.

VII. CONCLUSION

In this work, we present a MOO method for outlier detection and filtering of multimodal video datasets. By rigorously combining heterogeneous modalities—image, textual, and numerical data in particular—we show that each modality can be separately processed using specialized algorithms. For image data, we apply the NLM filter, which performs robust noise reduction while retaining crucial structural information. Conversely, text and numeric data are examined using the LOF algorithm, which detects and measures anomalies in terms of local density differences.

The effectiveness of the MOO framework is measured quantitatively by means of the SSIM and JSS, which are sound measures for quantifying the quality of the filtered data. Experimentally, our results show that MOO drastically improves the integrity of multimodal video data through efficient outlier elimination, thus enhancing the quality of the data as a whole. Such an approach is specially relevant for uses that demand high fidelity in video data, such as medical images, surveillance applications, and analysis of multimedia content. By meeting the challenges presented by multimodal data, MOO not just allows for enhanced outlier handling, but also safeguards against the very nature of multimodal data presenting diversity.

While the proposed Multimodal Outlier Optimization (MOO) method shows promising results in improving the quality of multimodal video data, there are a few limitations to consider. First, treating each modality separately though effective, may overlook interdependencies or correlations

between modalities that could provide richer context for identifying outliers. For instance, an anomaly that is subtle in one modality might only be clear when viewed alongside another. Second, the performance of the method depends on the quality and balance of the input data. In cases where one modality is noisy or incomplete (e.g., missing textual metadata), the filtering process might be less effective or biased. Additionally, the LOF algorithm can struggle with high-dimensional or highly sparse data, which is often the case in real-world multimodal datasets. Finally, although SSIM and JSS provide a solid quantitative measure of filtering effectiveness, they may not fully capture semantic or contextual nuances, especially in complex video content. Future work may benefit from exploring deeper integration between modalities and leveraging advanced models that can jointly learn from them.

This research enhances the development of multimodal data processing methods with an extensive solution improving the quality and usability of video data for use in diverse scientific and applied areas. Looking ahead, an important direction for future work is the integration of deep multimodal learning architectures. Unlike traditional approaches that treat each modality separately, deep models—such as transformers, multimodal autoencoders, or contrastive learning frameworks—can jointly learn representations across modalities, capturing complex relationships and context that may help in identifying more subtle or cross-modal outliers. Incorporating such models into the MOO framework could significantly improve its robustness and adaptability, especially in noisy or dynamic environments. Furthermore, exploring attention mechanisms within these architectures could enhance interpretability and allow the system to weigh the importance of different modalities depending on the context. These enhancements could make MOO suitable not just for pre-processing, but also for end-to-end applications in fields like healthcare diagnostics, smart surveillance, and autonomous navigation.

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests that could influence the work reported in this study.

DATA AVAILABILITY STATEMENT

The data utilized in this study has been synthetically generated using Python's Flaker Library. The data utilized for testing the performance of the proposition are publicly available and have been archived here.

REFERENCES

[1] M. Kuchroo, A. Godavarthi, A. Tong, G. Wolf, and S. Krishnaswamy, "Multimodal data visualization and denoising with integrated diffusion," in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2021, pp. 1–6, doi: 10.1109/MLSP52302.2021.9596214.

[2] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data Cognit. Comput.*, vol. 5, no. 1, p. 1, Dec. 2020, doi: 10.3390/bdcc5010001.

[3] B. K. S. Kumar, "Image denoising based on non-local means filter and its method noise thresholding," *Signal, Image Video Process.*, vol. 7, no. 6, pp. 1211–1227, Nov. 2013, doi: 10.1007/s11760-012-0389-y.

[4] M. Tang, Y. Kaymaz, B. L. Logeman, S. Eichhorn, Z. S. Liang, C. Dulac, and T. B. Sackton, "Evaluating single-cell cluster stability using the Jaccard similarity index," *Bioinformatics*, vol. 37, no. 15, pp. 2212–2214, Aug. 2021, doi: 10.1093/bioinformatics/btaa956.

[5] Y. Lin, L. Wan, S. Ma, and P. Zhang, "Feature structure similarity index for hybrid human and machine vision," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 1480–1484, doi: 10.1109/ICIP49359.2023.10222499.

[6] S. Kang and T. Park, "Detecting outlier behavior of game player players using multimodal physiology data," *Intell. Autom. Soft Comput.*, vol. 26, no. 1, pp. 205–214, 2019, doi: 10.31209/2019.100000141.

[7] A. Zlatintsi, P. P. Filntisis, N. Efthymiou, C. Garoufis, G. Retsinas, T. Sounapoglou, I. Maglogiannis, P. Tsanakas, N. Smyrnis, and P. Maragos, "Person identification and relapse detection from continuous recordings of biosignals challenge: Overview and results," *IEEE Open J. Signal Process.*, vol. 5, pp. 641–651, 2024, doi: 10.1109/OJSP.2024.3376300.

[8] N. E. H. Ben Chaabene, A. Bouzeghoub, R. Guetari, and H. H. B. Ghezala, "Deep learning methods for anomalies detection in social networks using multidimensional networks and multimodal data: A survey," *Multimedia Syst.*, vol. 28, no. 6, pp. 2133–2143, Dec. 2022, doi: 10.1007/s00530-020-00731-z.

[9] J. Wang, C. Wang, L. Guo, S. Zhao, D. Wang, S. Zhang, X. Zhao, J. Yu, Y. Wang, Y. Yang, S. Ma, and Q. Tian, "MDKAT: Multimodal decoupling with knowledge aggregation and transfer for video emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 10, pp. 9809–9822, Oct. 2025, doi: 10.1109/TCSVT.2025.3571534.

[10] C. Zhu, "Research on emotion recognition-based smart assistant system: Emotional intelligence and personalized services," *J. Syst. Manag. Sci.*, vol. 13, no. 5, pp. 227–242, 2023.

[11] R. Kannan, "Outlier detection for text data," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 1–12, doi: 10.1137/1.9781611974973.55.

[12] M. E. Eren, J. S. Moore, E. Skau, E. Moore, M. Bhattarai, G. Chennupati, and B. S. Alexandrov, "General-purpose unsupervised cyber anomaly detection via non-negative tensor factorization," *Digit. Threats, Res. Pract.*, vol. 4, no. 1, pp. 1–28, Mar. 2023, doi: 10.1145/3519602.

[13] D. Lee and K. Shin, "Robust factorization of real-world tensor streams with patterns, missing values, and outliers," in *Proc. IEEE 37th Int. Conf. Data Eng. (ICDE)*, Apr. 2021, pp. 840–851.

[14] J. Yin and J. Wang, "A model-based approach for text clustering with outlier detection," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, Helsinki, Finland, May 2016, pp. 625–636, doi: 10.1109/ICDE.2016.7498276.

[15] J. Kumar, S. U. Din, Q. Yang, R. Kumar, and J. Shao, "An online semantic-enhanced graphical model for evolving short text stream clustering," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13809–13820, Dec. 2022, doi: 10.1109/TCYB.2021.3108897.

[16] A. Interrante-Grant, M. Wang, L. Baer, R. Whelan, and T. Leek, "Synthetic datasets for program similarity research," 2024, *arXiv:2405.03478*.

[17] N. S. Zulkipli, S. Z. Satari, and W. N. S. Wan Yusoff, "A synthetic data generation procedure for univariate circular data with various outliers scenarios using Python programming language," *J. Phys., Conf. Ser.*, vol. 1988, no. 1, Jul. 2021, Art. no. 012111.

[18] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier detection: Methods, models, and classification," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–37, May 2021, doi: 10.1145/3381028.

[19] Z. Yuan, H. Chen, T. Li, X. Zhang, and B. Sang, "Multigranulation relative entropy-based mixed attribute outlier detection in neighborhood systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 8, pp. 5175–5187, Aug. 2022, doi: 10.1109/TSMC.2021.3119119.

[20] R. Mehmood and A. Kaur, "Modified difference squared image based non local means filter," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2020, pp. 1–7, doi: 10.1109/ICCCNT49239.2020.9225284.

[21] F. Sippel, J. Seiler, and A. Kaup, "Spatio-spectral image reconstruction using non-local filtering," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5, doi: 10.1109/VCIP53242.2021.9675421.

[22] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, and W. Gao, "Pretrained image processing transformer," 2021, *arXiv:2012.00364*.

[23] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," *Organizational Res. Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.

[24] C. P. Chai, "Comparison of text preprocessing methods," *Natural Lang. Eng.*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/s1351324922000213.
[25] E. A. Alshdaifat, D. A. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. D. F. S. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance," *Data*, vol. 6, no. 2, p. 11, 2021, doi: 10.3390/data6020011.
[26] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Syst. Appl.*, vol. 189, Mar. 2022, Art. no. 116087, doi: 10.1016/j.eswa.2021.116087.
[27] K. Papafitsoros and C. B. Schönlieb, "A combined first and second order variational approach for image reconstruction," *J. Math. Imag. Vis.*, vol. 48, no. 2, pp. 308–338, Feb. 2014.
[28] A. Bionda, L. Frittoli, and G. Boracchi, "Deep autoencoders for anomaly detection in textured images using CW-SSIM," in *Proc. Int. Conf. Image Anal. Process.*, May 2022, pp. 669-680, Cham, Switzerland: Springer, doi: 10.1007/978-3-031-06430-2_56.
[29] C. Wang, H. Zhu, J. Peng, Y. Wang, R. Yi, Y. Wu, L. Ma, and J. Zhang, "M3DM-NR: RGB-3D noisy-resistant industrial anomaly detection via multimodal denoising," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 11, pp. 9981–9993, Nov. 2025, doi: 10.1109/tpami.2025.3592089.
[30] R. N. A. Algburi, H. S. S. Aljibori, Z. Al-Huda, Y. H. Gu, and M. A. Al-Antari, "Advanced fault diagnosis in industrial robots through hierarchical hyper-Laplacian priors and singular spectrum analysis," *Complex Intell. Syst.*, vol. 11, no. 6, pp. 9981–9993, Jun. 2025.
[31] R. N. Ali Algburi and H. Gao, "Detecting feeble position oscillations from rotary encoder signal in an industrial robot via singular spectrum analysis," *IET Sci., Meas. Technol.*, vol. 14, no. 5, pp. 600–609, Jul. 2020.

KRITTIKA DAS is currently pursuing the B.Tech. degree in computer science and engineering with Techno International New Town, Kolkata, India. She has contributed to several journal articles and conference papers. Her research interests include deep learning, AI/ML, web technology, and natural language processing (NLP).

NILANJAN DEY (Senior Member, IEEE) received the B.Tech. and M.Tech. degrees in information technology from West Bengal Board of Technical University, in 2005 and 2011, respectively, and the Ph.D. degree in electronics and telecommunication engineering from Jadavpur University, Kolkata, India, in 2015. Currently, he is a Professor with the Techno International New Town, Kolkata, and a Visiting Fellow with the University of Reading, U.K. He has authored more than 300 research papers in peer-reviewed journals and international conferences and 40 books. His research interests include medical imaging and machine learning. He is also the Editor-in-Chief of *International Journal of Ambient Computing and Intelligence*, an Associate Editor of IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY, and a series Co-Editor of *Springer Tracts in Nature-Inspired Computing* and *Data-Intensive Research* from Springer Nature and *Advances in Ubiquitous Sensing Applications for Healthcare* from Elsevier.

BITAN MISRA (Member, IEEE) received the B.Tech. and M.Tech. degrees in electronics and telecommunication engineering from KIIT University, Bhubaneswar, India, in 2018, and the Ph.D. degree from the National Institute of Technology, Durgapur, India, in 2022. She has published several research papers in various international journals, book chapters and conferences, and five authored books. She holds multiple patents and copyrights. Her main research interests include optimization techniques, deep learning, evolutionary algorithms, and soft computing techniques. She has worked as a reviewer on several national and international journals and conferences. She is also an Associate Editor of *International Journal of Ambient Computing and Intelligence*.

SATYABRATA ROY (Senior Member, IEEE) received the B.Tech., M.Tech., and Ph.D. degrees in computer science and engineering, in 2009, 2014, and 2020, respectively. He is currently an Associate Professor with the Department of Computer Science and Engineering, School of Computer Science and Engineering, Manipal University Jaipur, Rajasthan, India. He is an enthusiastic and motivating technocrat with more than ten years of research and academic experience at different reputed institutes. He has supervised many students for their M.Tech. dissertation work and supervising the Ph.D. scholars with Manipal University Jaipur. He has published many research papers in top-quality international journals and national/international conferences of repute. His research interests include cryptography, the Internet of Things, cellular automata, computer networks, computational intelligence, machine learning, and formal languages.

R. SIMON SHERRATT (Fellow, IEEE) received the B.Eng. degree from Sheffield City Polytechnic, in 1992, and the M.Sc. and Ph.D. degrees from the University of Salford, in 1993 and 1996, respectively. In 1996, he was appointed as a Lecturer in electronic engineering with the University of Reading, where he is currently a Professor of biosensors. His research area is wearable devices, mainly for healthcare and emotion detection. He was awarded the First Place IEEE Chester Sall Award, in 2004, the Second Place, in 2014 and 2024, the Third Place, in 2015, and the Third Place, in 2016, for best papers in IEEE TRANSACTIONS ON CONSUMER ELECTRONICS.