TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Frassini, E., Butler, R. M., Hendriks, B. H. W., van der Elst, M., & van den Dobbelsteen, J. J. (2026). Unsupervised multi-modal variational autoencoder for anomaly detection in coronary angiography. *Minimally Invasive Therapy and Allied Technologies*. https://doi.org/10.1080/13645706.2026.2705585

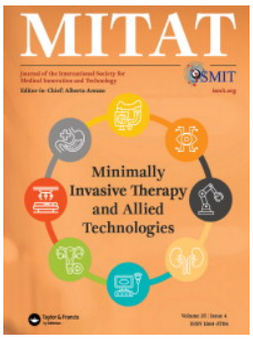
Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.



Taylor & Francis
Taylor & Francis Group

Minimally Invasive Therapy & Allied Technologies

ISSN: 1364-5706 (Print) 1365-2931 (Online) Journal homepage: www.tandfonline.com/journals/imit20

Unsupervised multi-modal variational autoencoder for anomaly detection in coronary angiography

Emanuele Frassini , Rick M. Butler , Benno H. W. Hendriks , Maarten van der Elst & John J. van den Dobbelsteen

To cite this article: Emanuele Frassini , Rick M. Butler , Benno H. W. Hendriks , Maarten van der Elst & John J. van den Dobbelsteen (23 Jul 2026): Unsupervised multi-modal variational autoencoder for anomaly detection in coronary angiography, Minimally Invasive Therapy & Allied Technologies, DOI: 10.1080/13645706.2026.2705585

To link to this article: https://doi.org/10.1080/13645706.2026.2705585

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 23 Jul 2026.

Submit your article to this journal

Article views: 196

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=imit20

MINIMALLY INVASIVE THERAPY & ALLIED TECHNOLOGIES
https://doi.org/10.1080/13645706.2026.2705585

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

Unsupervised multi-modal variational autoencoder for anomaly detection in coronary angiography

Emanuele Frassini[a], Rick M. Butler[a], Benno H. W. Hendriks[a,b], Maarten van der Elst[a,c] and John J. van den Dobbelsteen[a]

[a]Delft University of Technology, Delft, The Netherlands; [b]Philips Medical Systems, Best, The Netherlands; [c]Reinier de Graaf Hospital, Delft, The Netherlands

ABSTRACT

Background: Coronary angiography (CAG) is among the most common procedures in catheterization laboratories. Patient-specific vessel tortuosity may require additional catheter exchanges, increasing procedure duration and complication risk. Detecting such workflow disruptions automatically is challenging because anomalies are subtle and difficult to capture using a single data modality.

Methods: We propose a novel unsupervised multi-modal anomaly detection framework based on a variational autoencoder (VAE). Video and pose data from 191 real-world CAG procedures were segmented into 10-second clips. Features extracted with 3D convolutional networks were fused through modality-specific LSTM encoders. Reconstruction error (mean squared error) was used as the anomaly score, and thresholds were calibrated on a held-out dataset.

Results: Procedures requiring additional catheter exchanges accounted for 29.8% of the dataset and had a 34% longer mean duration than normal procedures (47.6 vs. 35.5 min). The multi-modal VAE outperformed unimodal models, achieving an F1-score of 0.83, AUC-ROC of 0.88, and AUC-PR of 0.90. Detection performance peaked during the period when anomalies most frequently occurred.

Conclusions: This study demonstrates the effectiveness of multi-modal fusion for unsupervised anomaly detection in complex surgical workflows. These findings support the development of intelligent workflow monitoring systems for catheterization laboratories.

ARTICLE HISTORY
Received 26 November 2025
Accepted 14 June 2026

KEYWORDS
Catheterization laboratory; coronary angiography; anomaly detection; unsupervised learning; variational autoencoder

Introduction

In the catheterization laboratory (cath lab) a variety of cardiac procedures are performed. Coronary angiography (CAG) is one of the most common cardiac procedures and consists of standardized steps, such as catheter insertion and X-ray acquisitions [1]. This standardization makes CAG an ideal procedure for developing intelligent systems that can automatically monitor and understand the clinical workflow. However, even the workflow of routine procedures is never fully predictable as patient-specific anatomical and pathological factors—such as vessel tortuosity, stenoses, calcification, vessel narrowing, and dissections—can affect procedural complexity. Cardiologists may sometimes face challenges in navigating the initially selected catheter, necessitating the use of an alternative one with different shape or stiffness [2], which can potentially lead to increased procedure times and other complications [3].

Intelligent systems designed to support clinicians perform tasks effectively require a deep understanding of the clinical workflow [4]. The ability to accurately and automatically detect disruptions and adaptations during the procedures is important to enhance workflow efficiency and to ensure high-quality clinical outcomes [5]. However, this task presents a unique set of challenges. In dynamic clinical environments, where instruments, patients, and personnel interact in complex ways [5], events such as catheter exchanges are often subtle and multifaceted, manifesting only slightly in the visual appearance of the scene as well as in the kinematics and spatiotemporal relationships between the actions of the cardiologist [6].

CONTACT Emanuele Frassini ema.frassini@hotmail.com Delft University of Technology, Mekelweg 2, 2628CD, Delft, The Netherlands

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

In this paper we approach these detection challenges by considering the disruptions of the standard workflow as procedural anomalies. Anomaly detection is a well-established field, with methods ranging from statistical process control to more advanced deep learning techniques like long short term memory network (LSTM) and autoencoders [6]. More recently, pose-based action recognition and anomaly detection have gained traction by leveraging kinematic data from human skeletons to model actions and identify deviations from normal behavior [7]. However, these unimodal approaches often fail to capture the holistic context of complex events [8]. Approaches that depend solely on video streams may struggle to identify subtle kinematic errors or procedural deviations that are not visually distinct, while methods relying on kinematics alone lack contextual information about the operational environment [9]. The field of multi-modal fusion has explored strategies for combining information from disparate sources, though applications for anomaly detection in the clinical field are limited [10]. Finally, variational autoencoders (VAEs) have proven to be a powerful unsupervised method for anomaly detection by learning a latent representation of normal data and using reconstruction error as an anomaly score [11]. Despite these advances, such approaches have yet to be applied to the analysis of procedures performed in catheterization laboratories, where the complexity of clinical workflows and the criticality of timely anomaly detection make this an especially compelling and underexplored application domain [12].

This paper introduces an unsupervised multi-modal VAE framework for anomaly detection in cath lab procedures. High-dimensional features are extracted from video and pose data using 3D convolutional models originally developed for action recognition, with the final classification layer removed. Videos are segmented into 10-s clips, and features from each segment are sequentially combined into a unified representation of the entire procedure before being input to the VAE. The reconstruction error of the VAE is then used as the primary score to distinguish between normal and anomalous procedures. The reconstruction error at each clip is also investigated to better understand the behavior of the model. The framework is validated on a real-world medical dataset.

Material and methods

Our dataset consists of video recordings and pose data from 191 CAG procedures performed at Reinier de Graaf Hospital in Delft, the Netherlands, and described in detail in [13–15]. We analyzed the images captured from the south corner of the cath lab (Figure 2), as it provides a clear yet narrow focus on the operating table surroundings, limiting occlusions from reflections and external elements [16]. We define anomalies as the occurrence of an additional catheter change within the procedure.

Framework

Our proposed framework is an end-to-end pipeline designed to detect anomalies in video procedures. Figure 1 shows the overview of our framework.

Each video procedure is segmented into 10-s clips (Figure 1A). For each segment, we first perform feature extraction using pre-trained models to obtain high-dimensional representations from two distinct modalities: RGB and poses (Figure 1B). RGB video data are used to capture visual context, whereas pose data can capture kinematic motion. These features are then concatenated (Figure 1C) and fed into our core component, the multi-modal VAE (Figure 1D). This network employs an intermediate fusion strategy, where features from each modality are first encoded separately by dedicated LSTM layers before being concatenated and

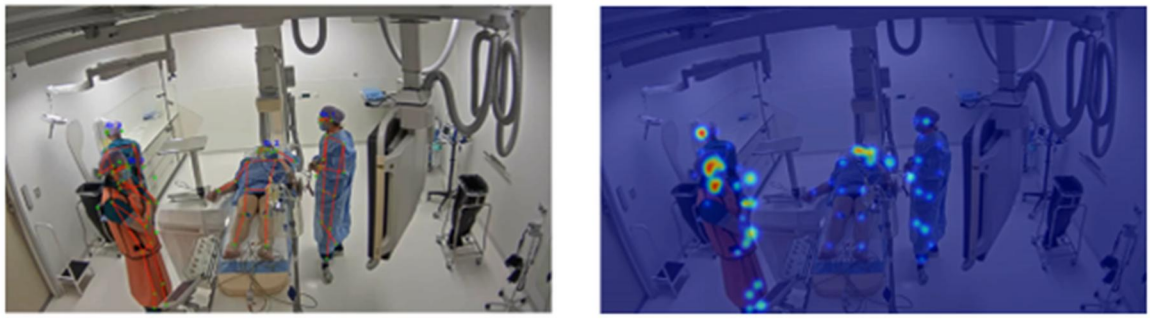


Figure 1. Example of video recording of a CAG procedure with poses (left) and heatmaps (right).

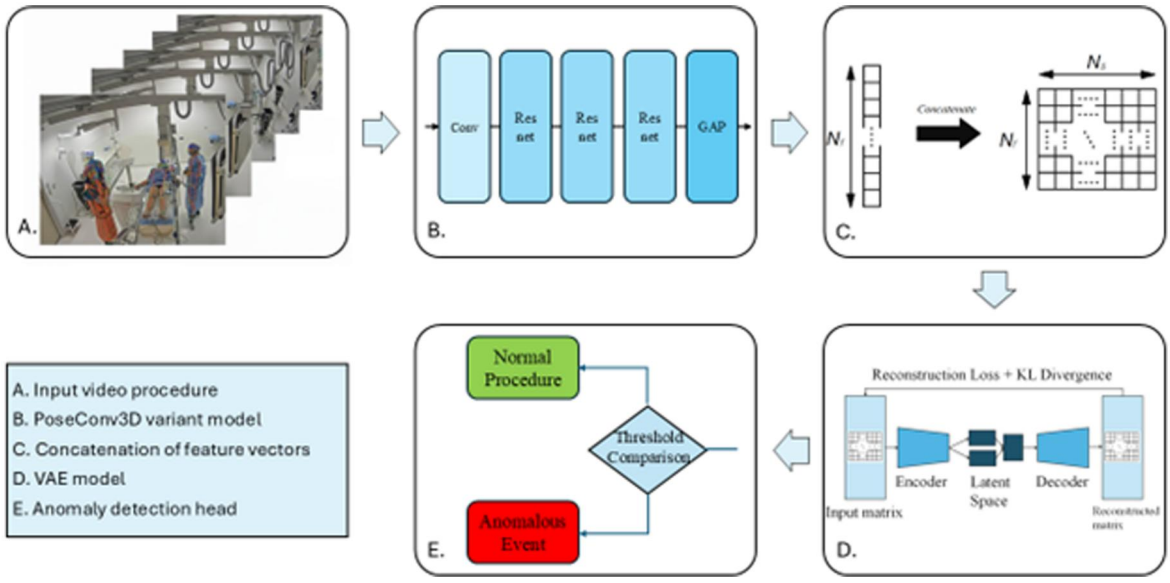


Figure 2. Overview of the proposed unsupervised multi-modal anomaly detection framework. (A) The input video is segmented into 10-s clips; (B) Each clip is processed by a pre-trained PoseConv3D network to extract high-dimensional features; (C) The feature vectors from consecutive clips are concatenated to form a single matrix representing the entire surgical procedure; (D) This feature matrix is then fed into a Variational Autoencoder, which learns to reconstruct the input; (E) The anomaly score is calculated based on the reconstruction error.

fed into the latent space of the VAE. During the anomaly detection phase, the framework uses the trained VAE to compute an anomaly score for a given procedure based on its reconstruction error. The score is given by the mean squared error (MSE) between the original input sequence and its VAE-reconstructed version. The error represents how poorly the model can reconstruct the input. Finally, this score is compared against a predefined threshold, which is calibrated on a held-out dataset, to classify the procedure as either normal or anomalous (Figure 1E).

The proposed unsupervised anomaly detection approach was implemented in Python, and the source code is publicly available at: httpS://GITHUB.COM/EMANUELEFRASSINI/UNSUPERVISED_ANOMALY_DETECTION.GIT.

Feature extraction

The process began with the extraction of 2D poses from the video frames using a top-down pose tracker [13]. Each pose file was divided into segments of 10 s. The 2D poses were converted into a 3D heatmap volume by stacking heatmaps of joints and limbs along the temporal dimension. The joint heatmaps were obtained by composing gaussian maps centered at every joint, whereas a distance function was used to the segment between two joints to create a limb heatmap along the temporal dimension [14]. This volume was then fed into a pre-trained PoseConv3D network, a model based on a three-dimensional convolutional neural network (3D CNN) for skeleton-based action recognition, which served as the backbone for feature extraction [14]. We removed the final layer of the PoseConv3D model to obtain the embeddings of the last layer. From this network, we extracted 512-dimensional embeddings from the penultimate layer, originally intended for the classification head, yielding one feature vector per 10-s clip of each CAG procedure.

Figure 2 shows an example video recording from the south wall, illustrating both the extracted poses and the corresponding heatmaps.

In addition to kinematic pose data, our framework incorporates a separate stream for RGB video features. RGB videos are segmented into 10-s clips and processed using a modified RGBPoseConv3D network for raw RGB frames [14]. This 3D CNN backbone extracted 2048-dimensional feature vectors for each video segment.

The feature vectors, with N_f features, obtained through this process are sequentially concatenated into a matrix with dimensions $N_s \times N_f$, until the end of each procedure. A padding strategy is employed to handle procedures of different durations, where the sequence is filled with zeros to meet the required length N_s. This ensures a uniform input size for the LSTM layers. Finally, all extracted features are normalized over the feature axis to a [0,1] range.

Multi-modal VAE architecture

Our multi-modal VAE architecture is comprised of three primary components: an encoder network, a latent space, and a decoder network. The encoder network consists of two separate LSTM layers, each dedicated to a single modality. One LSTM processes the pose feature sequences with an input dimension of 512, while the other processes the RGB video feature sequences with an input dimension of 2048. Each final hidden state of the LSTM layers is extracted. Furthermore, we employ an intermediate fusion strategy where these two hidden states are concatenated to form a single, unified representation of the input. This fused vector is then passed through two separate fully connected layers to generate the parameters for the latent space distribution: the mean and log-variance. In the latent space, a single latent vector is sampled from this learned Gaussian distribution using the reparameterization trick, which allows gradients to flow back through the sampling process during training [15]. The decoder network takes this sampled latent vector, expands it, and feeds it into a single LSTM layer. The output of this LSTM is then passed through a final fully connected layer to reconstruct the original combined multi-modal input vector.

Experimental setup

Each procedure video was divided into 10-s segments. Features extracted from these segments were concatenated to form a matrix representing the entire procedure, which was then input to the VAE model. The training of our multi-modal VAE is driven by a single, comprehensive loss function:

$$L_{VAE} = L_{REC} + \beta L_{KL}.$$

The first component L_{REC} measures how accurately the model can reconstruct the original input features from the learned latent representation. It is calculated as the sum of the MSE losses for both the pose and RGB components, comparing the reconstructed output to the original, normalized input features. The second component is the Kullback–Leibler divergence loss (L_{KL}). This term acts as a regularizer, encouraging the latent distribution learned by the encoder to be close to a standard normal distribution [17]. The hyperparameter β serves as a weighting factor. It balances these two competing objectives: a low β prioritizes accurate reconstruction of the input data, while a high β forces the model to learn a more well-structured, regularized latent space [18].

The dataset was first divided into normal and anomalous procedures. Each subset was further partitioned into training, validation, calibration, and test sets, with 50% used for training the VAE, 10% for validation, and the remaining 40% evenly split between calibration and testing. We evaluate our anomaly detection framework using standard metrics, with a focus on the F1-score due to the imbalanced nature of our dataset. Precision and recall assess the ability of the model to minimize false positives and missed anomalies, respectively, and the F1-score balances these objectives. Additionally, we report the area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision-recall curve (AUC-PR), the latter providing a robust measure of performance for rare events.

The core of our anomaly detection strategy relies on the ability of the VAE to model 'normal' data and identify deviations. We define the anomaly score for any given input sequence as the overall reconstruction error, calculated as the MSE between the original combined multi-modal sequence and the sequence reconstructed by the VAE. The rationale is that for normal data, which the VAE was trained on, the reconstruction error will be low. In contrast, an anomalous sequence will be difficult for the model to reconstruct. This difficulty arises because anomalous events often exhibit distinct patterns in pose kinematics and RGB features, such as abrupt movements, or unexpected visual elements that deviate from the statistical regularities learned during training. Neural networks, including VAEs, generally interpolate well within the training data but struggle to extrapolate beyond it, which is evident when out-of-distribution inputs, such as anomalies, cause the network to lack the learned representations for accurate reconstruction, leading to higher reconstruction errors [19]. To classify a sequence as anomalous or normal, we apply a thresholding strategy. An optimal threshold is determined by using a separate calibration set that includes both normal and known anomaly samples. We iterate through a range of potential threshold values and select the one that maximizes the F1-score.

Results

Our dataset consists of 134 normal and 57 anomalous procedures (29.8%). We began by investigating the clinical relevance of our framework. We analyzed the impact of the additional change of catheter, identified as the detected anomaly, on the overall duration of the procedure. Figure 3 presents a kernel density

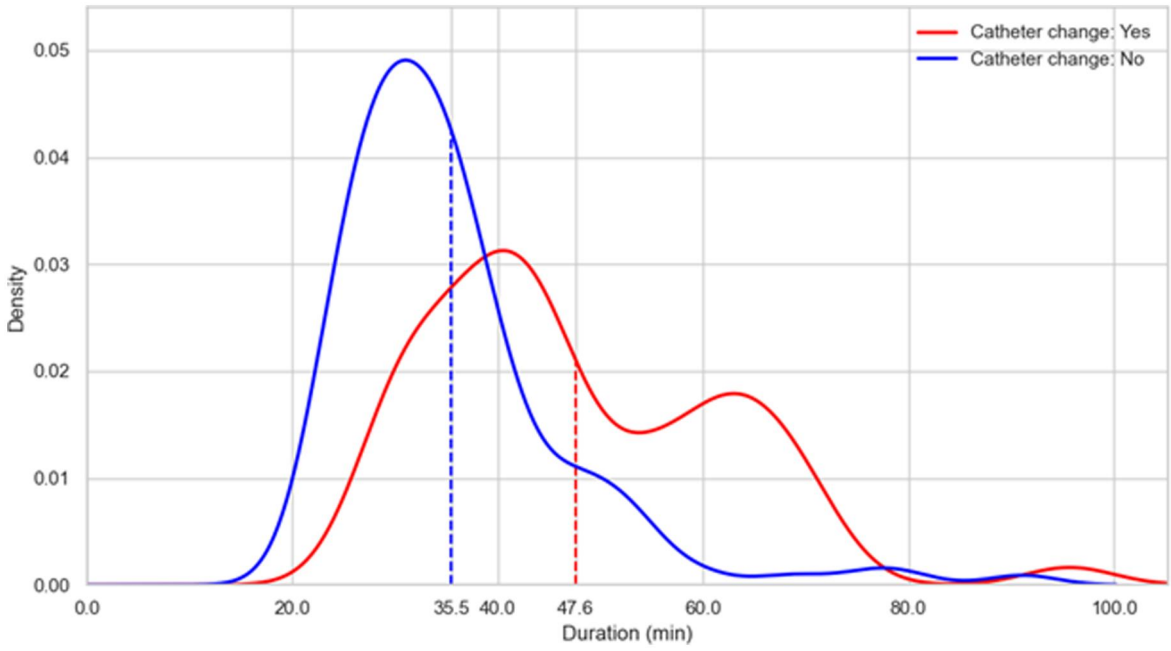


Figure 3. Distributions of the duration of a CAG procedure with (red line) and without (blue line) the additional change of catheter.

estimation of procedure durations for both normal cases and those with the anomaly. As shown, the distribution for procedures with the anomaly (red line) is shifted to the right, indicating a longer overall duration. The mean duration for procedures without the anomaly was 35.5 min, while procedures containing the additional catheter change had a mean duration of 47.6 min. This difference represents an increase of approximately 34%. We performed a Monte Carlo simulation using the Mann–Whitney U test across 100,000 runs to validate the statistical significance of this difference. The analysis yielded a p-value of 0.00023 ± 0.0027 ($p < 0.001$).

Quantitative performance

The performance of our proposed framework was evaluated across three distinct modalities: pose features, RGB video features, and a fused combination of both. Figure 4 depicts a heatmap of the highest performance metrics we achieved. The unimodal pose-based model achieved an F1-score of 0.70, an AUC-PR of 0.71 and an AUC-ROC of 0.68. However it can be noticed that the recall stands at 90%, much higher than the precision (58%). The video-based model demonstrated superior performance, with an F1-score of 0.81, an AUC-PR of 0.80 and an AUC-ROC of 0.86. The highest performance was consistently achieved by the multi-modal fusion model, which combines both pose and video features. This approach yielded an F1-score of 0.83, an AUC-ROC of 0.88, and an AUC-PR of 0.90.

A detailed analysis of the performance of the models on the validation set is provided in the confusion matrix component breakdown in Figure 5. The figure illustrates the counts of true positives (Tp), false positives (Fp), false negatives (Fn), and true negatives (Tn) for each of the three modalities. The pose-only model shows a high count of both true positives and false positives, indicating its tendency to flag a large number of events as anomalous (positive), whether they are true anomalies or not. In contrast, the other two modalities demonstrate a notable reduction in false cases (orange and green). The model that used video and poses together showed a slight reduction in false cases compared with the video-only model.

Figure 6 illustrates the average cumulative accuracy of the model in detecting anomalies at various points during a procedure, compared with the actual distribution of when these anomalies occur. The lower plot shows that most anomalies tend to cluster around the 35 to 45-min mark of the procedure. The upper plot exhibits a visual correlation with this distribution. Accuracy is relatively low at the beginning of the procedure, below 60%. However this value rises sharply to over 80% as the procedure enters the time window where anomalies are most common.

Discussion

Our study introduces a novel unsupervised framework for anomaly detection, demonstrating the power of a

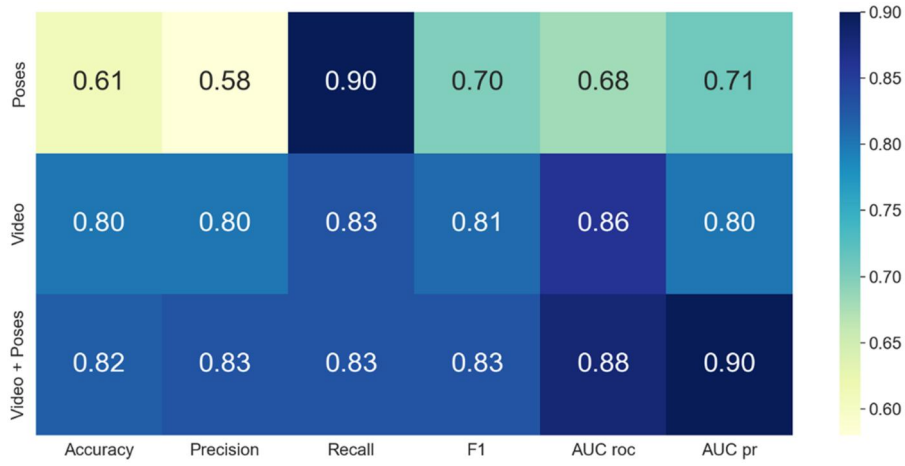


Figure 4. Heatmap of the performance metrics of our anomaly detector with different input data.

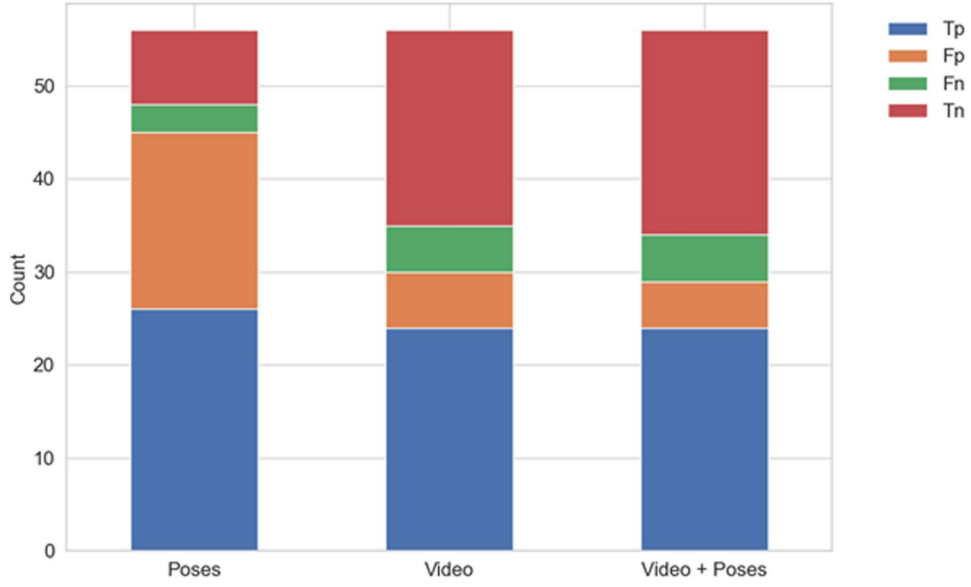


Figure 5. Confusion matrix results of the anomaly detection model. Columns represent different input modalities: pose data only (left), video data only (middle), and a combination of both (right).

multi-modal VAE approach in the cath lab setting. The fusion of kinematic data from pose estimation with rich visual context from video outperforms unimodal counterparts, achieving the highest F1-score of 0.83, AUC-ROC of 0.88, and AUC-PR of 0.90. This finding confirms our hypothesis that a holistic understanding of surgical procedures, incorporating both the visual context and the kinematic motion, can accurately identify procedural deviations.

A key insight from our results is the performance gap between the unimodal models. The video-based model, with an F1-score of 0.81, outperformed the pose-based model's F1-score of 0.70. This suggests that the visual context, captured by the PoseConv3D network on raw RGB frames, contains discriminative information crucial for anomaly detection. While the pose-based model showed a high recall (90%), it struggled with a lower precision (58%). This indicates a tendency to flag many normal events as anomalous and highlights the inherent ambiguity of isolated kinematic movements. A movement of the cardiologist might appear unusual on its own, but when placed in the full context of the procedure, it can be identified as a normal action. The multi-modal fusion framework successfully mitigates this weakness by leveraging the video stream to provide this critical context, resulting in a substantial reduction in false positives and a more robust and reliable detection system.

Furthermore, our temporal analysis provides a crucial validation of the clinical relevance of our approach. The significant increase in the cumulative accuracy (Figure 6) at the exact time window where most anomalies naturally

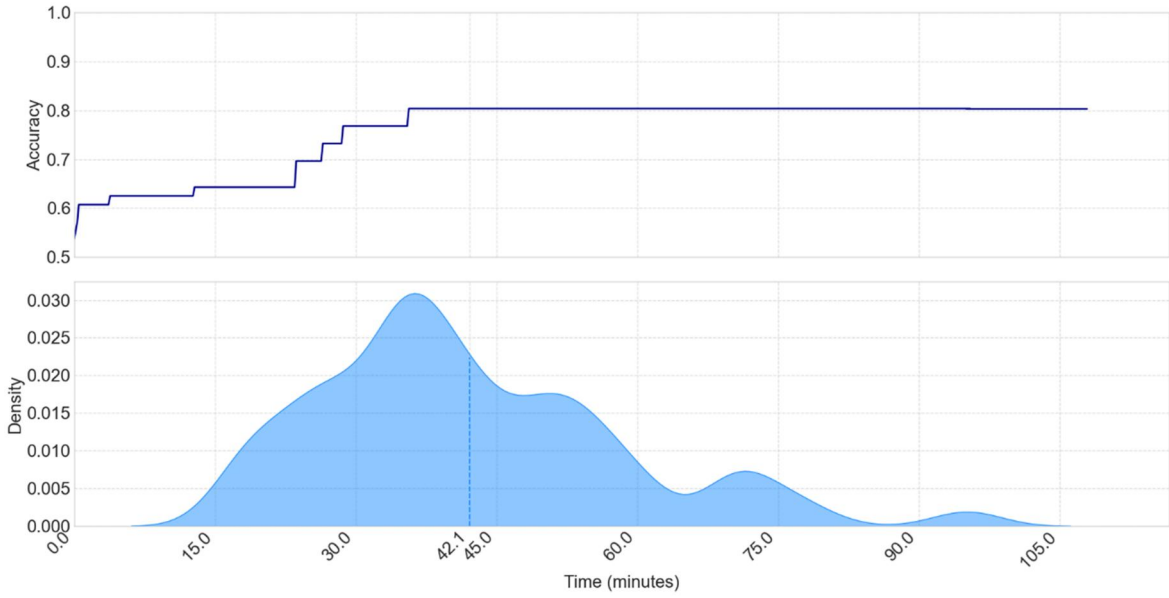


Figure 6. Average cumulative accuracy of the model (upper plot) compared with the temporal distribution of anomalies (lower plot) in CAG procedures.

occur is a compelling result. This finding not only validates our model but also points to its potential for real-world clinical application. An effective anomaly detection system is most valuable when it provides a timely warning. The sharp rise in accuracy from below 60% to over 80% during the peak anomaly period suggests that our framework is not simply detecting a difference from the training data, but is learning the underlying temporal progression of the procedure and flagging deviations precisely when they are most likely to have clinical significance. The fact that the detected anomalies are associated with a significant increase in procedure duration (approximately 34%) further underscores the clinical impact of such events.

The current study was limited to CAG procedures. However, as the prevalence of tortuosity is common, especially in older patients or those with atherosclerosis, adaptations of the workflow frequently occur in all types of endovascular procedures. Many workflows (neuro, coronary, peripheral) are designed with this in mind, and standard setups usually include multiple catheter types, because tortuosity is expected. Still, while expected in principle, the degree and exact shape of tortuosity often surprise operators. Severe tortuosity can force a switch in access route, catheter type, or escalation of catheter changes, with tremendous impact on time and safety aspects: it slows the procedure, increases radiation and contrast exposure, and raises complication risk (dissection, vasospasm). An effective Intelligent system that detects necessary adaptations to the workflow can help streamlining and preparing subsequent procedure steps to reduce the impact and risks as much as possible [20].

Additionally, while the cath lab environment naturally contains routine background noise such as staff interactions, the training of the model on normal procedures allows it to characterize these common events as part of the baseline procedural distribution. This robustness is reflected in our low false-positive rate, suggesting that the model successfully differentiates non-clinical background fluctuations from the targeted anomalies.

This study opens avenues for future research. A key limitation, common to unsupervised methods, is the assumption that anomalies yield higher reconstruction errors. While this holds true for our findings, some anomalies may be well reconstructed or resemble normal data. Future work could improve discrimination by adopting a dual-objective training strategy: minimizing reconstruction error for normal data while maximizing it for a small set of known anomalies. This contrastive learning approach would encourage the VAE to learn more distinct latent representations for normal and anomalous inputs.

In conclusion, we have presented an effective multi-modal framework for unsupervised anomaly detection in complex surgical workflows. Our multi-modal VAE demonstrated that combining visual and kinematic features provides a more comprehensive and accurate understanding of CAG procedures than either modality alone. The use of an unsupervised method is particularly valuable given the scarcity and diversity of labeled anomalies in real-world clinical datasets. This work is an important step toward building intelligent systems that can support cardiologists in the operating room, and enhance efficiency and patient safety.

Acknowledgments

We would like to thank Silvia Pintea and Vincent van der Sluis for their valuable input in designing the model architecture, and Jan Constandse for his guidance in defining the clinical relevance of this work.

Author contributions

CRediT: **Emanuele Frassini**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Rick M. Butler**: Conceptualization, Formal analysis, Software; **Benno H. W. Hendriks**: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing; **Maarten van der Elst**: Conceptualization, Supervision, Writing – review & editing; **John J. van den Dobbelsteen**: Conceptualization, Methodology, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Manda YR, Baradhi KM. Cardiac catheterization risks and complications. In StatPearls [Internet]. Treasure Island (FL): statPearls Publishing; 2023. Available from: httpS://www.NCBI.NLM.NIH.GOV/BOOKS/NBK531461/

[2] Ali A, Dodou D, Smit G, et al. Stabilizing interventional instruments in the cardiovascular system: a classification of mechanisms. Med Eng Phys. 2021; 89(1):22–32. doi: 10.1016/j.medengphy.2021.01.004.

[3] Frassini E, Vijfvinkel TS, Butler RM, et al. Deep learning methods for clinical workflow phase-based prediction of procedure duration: a benchmark study. Comput Assist Surg (Abingdon). 2025;30(1): 2466426. doi: 10.1080/24699322.2025.2466426.

[4] Padoy N. Machine and deep learning for workflow recognition during surgery. Minim Invasive Ther Allied Technol. 2019;28(2):82–90. doi: 10.1080/13645706.2019.1584116.

[5] AlDhaen FS. AI-powered transformation of healthcare: enhancing patient safety through AI interventions with the mediating role of operational efficiency and moderating role of digital competence-insights from the gulf cooperation council region. Healthcare (Basel). 2025;13(6):614. doi: 10.3390/healthcare13060614.

[6] Zamanzadeh Darban Z, Webb GI, Pan S, et al. Deep learning for time series anomaly detection: a survey. ACM Comput Surv. 2024;57(1):1–42. doi: 10.1145/3691338.

[7] Noghre GA, Pazho AD, Katariya V, et al. Understanding the challenges and opportunities of pose-based anomaly detection. In Proceedings of the 8th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence (iWOAR '23). 2023. Sep 21-22; Lübeck, Germany. New York (NY): association for Computing Machinery; 2023. p. 9.

[8] Boretto L, Pelanis E, Regensburger A, et al. Hybrid optical-vision tracking in laparoscopy: accuracy of navigation and ultrasound reconstruction. Minim Invasive Ther Allied Technol. 2024;33(3):176–183. doi: 10.1080/13645706.2024.2313032.

[9] Zappella L, Béjar B, Hager G, et al. Surgical gesture classification from video and kinematic data. Med Image Anal. 2013;17(7):732–745. doi: 10.1016/j.media.2013.04.007.

[10] Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects. Proc IEEE. 2015;103(9):1449–1477. doi: 10.1109/JPROC.2015.2460697.

[11] Nguyen HH, Nguyen CN, Dao XT, et al. Variational autoencoder for anomaly detection: a comparative study. arXiv e-Prints. 2024; 2408.13561. Accepted for poster presentation at IEEE ICCE 2024.

[12] Henien S, Aronow HD, Abbott JD. Quality management in the cardiac catheterization laboratory. J Thorac Dis. 2020;12(4):1695–1705. doi: 10.21037/jtd.2019.12.81.

[13] Butler RM, Vijfvinkel TS, Frassini E, et al. 2D human pose tracking in the cardiac catheterisation laboratory with BYTE. Med Eng Phys. 2025;135(1): 104270. doi: 10.1016/j.medengphy.2024.104270.

[14] Duan H, Zhao Y, Chen K, et al. Revisiting skeleton-based action recognition. arXiv e-Prints. 2022; 2104.13586. CVPR 2022 (Oral Presentation).

[15] Kingma DP, Salimans T, Welling M. Variational dropout and the local reparameterization trick. Advances in Neural Information Processing Systems. 2015; 28:2575–2583.

[16] Butler RM, Frassini E, Vijfvinkel TS, et al. Benchmarking 2D human pose estimators and trackers for workflow analysis in the cardiac catheterization laboratory. Med Eng Phys. 2025;136(1):104289. doi: 10.1016/j.medengphy.2025.104289.

[17] Cui J, Zhu B, Xu Q, et al. Generalized Kullback-Leibler divergence loss. IEEE Trans Pattern Anal Mach Intell. 2026;1–12. doi:10.1109/TPAMI.2026.3703828.

[18] Asperti A, Trentin M. Balancing reconstruction error and Kullback-Leibler divergence in variational autoencoders. IEEE Access. 2020;8:199440–199448. doi: 10.1109/ACCESS.2020.3034828.

[19] Zhan J, Xie X, Mao J, et al. Evaluating interpolation and extrapolation performance of neural retrieval models. In. proceedings of the 31st ACM international conference on information and knowledge management (CIKM '22). New York (NY): ACM; 2022. p. 2486–2496 doi: 10.1145/3511808.3557312.

[20] Pescio M, Kundrat D, Dagnino G. Endovascular robotics: technical advances and future directions. Minim Invasive Ther Allied Technol. 2025;34(3): 239–252. doi: 10.1080/13645706.2025.2454237.