Explainable Artificial Intelligence for Event and Anomaly Detection in Healthcare Monitoring Systems

Ayush Amol Deshmukh[1], Sangharsh Suresh Ghoble[2], Dr. Vandana Nilesh Pagar[3]
[1,2]*Department of Electronics & Computer Engineering, MIT Arts, Commerce and Science College, Alandi, Pune, Maharashtra, India*
[3]*Mentor, Department of Electronics & Computer Engineering, MIT Arts, Commerce and Science College, Alandi, Pune, Maharashtra, India*
doi.org/10.64643/ijirt.208483-459

Abstract*—The growing uses of artificial intelligence (AI) in the form of beside vital-sign monitors, wearable sensors or Internet of Medical Things (IoMT) devices allows early identification of critical events and various anomalies including arrhythmias, sepsis onset, falls, or physiological deterioration. The underlying deep learning and ensemble models, however, are typically opaque, undermining confidence in their clinical recommendations, regulatory approval, and patient safety in the case of an unsubstantiated event detection. Explainable Artificial Intelligence (XAI) aims to address this challenge by providing justifications and explanations accompanying the model's predictions that are understandable to a human. The paper provides an overview of the state-of-the-art in XAI for events and anomalies detection in healthcare monitoring and discusses opportunities and barriers to adopting XAI methods in the application domain. We summarize the main explanation approaches, provide an overview of their use in the context of electrocardiogram (ECG) analysis, IoMT security, behavioural healthcare monitoring and ICU deterioration prediction, and articulate a general conceptual framework. The paper concludes with a discussion of open challenges and future research direction.

Index Terms*—Explainable Artificial Intelligence, XAI, anomaly detection, event detection, healthcare monitoring, IoMT, SHAP, LIME, Grad-CAM, interpretability, clinical decision support.

I. INTRODUCTION

Healthcare monitoring systems including bedside patient monitors, remote patient monitoring (RPM) solutions, wearable biosensors and IoMT continuously generate vast amounts of data including ECG, photoplethysmography (PPG), respiratory rate, oxygen saturation and other physiologic waveforms and vital signs to support detection of clinically relevant events and anomalies. These range from cardiac arrhythmias and sepsis to falls, apnea, and device-level attacks among others. Event detection in healthcare monitoring is typically addressed using machine learning and deep learning (DL) methods including convolutional, recurrent networks, or autoencoders.

A current limitation of current approaches, however, is that the internal working of these models is typically opaque, often undermining confidence in their predictions at the clinical decision-making frontline. A medical professional presented with a model alert about anomalous events may well ask - What objective justification supports this conclusion? The ability to formally justify or explain a model prediction is becoming an important requirement to ensure clinical acceptance of these systems. Explainable Artificial intelligence (XAI) aims to address this challenge by providing human-facing justifications, feature attributions, saliency amps or counterfactuals that accompany the models' predictions without necessarily sacrificing model performance 1,2

The paper presents an overview of relevant work in the area of XAI and details how different explanation approaches are used to support event and anomaly detection in healthcare monitoring. The specific contributions of this paper and the structure are as follows:

1. Section III reviews the fundamental explanation approaches.
2. Section IV describes key aspects of anomaly and event detection from the perspective of healthcare monitoring.

3. Section V articulates a conceptual framework that integrates the explanation generation and visualization components into the event detection pipeline.
4. Section VI discusses applications and case studies.
5. Section VII addresses the evaluation and ranking of different approaches.
6. Section VIII outlines key challenges and opportunities and identifies directions for future research.

II. RELATED WORK

An early survey by Adadi and Berrada provides an overview of different XAI methods and categorizes them according to the scope (local versus global), the model dependency (model-agnostic versus model-specific), and form of presentation among criteria. Mienye et al. offer an overview of the application of XAI methods in healthcare and discuss considerations relating to regulatory approvals (e.g., EU Act) and data privacy (HIPPA, GDPR). They observed that interpretability often comes at the expense of model performance. Using a similar theme, Allgaier et al. describe a systematic review of how ML models produce predictions to satisfy clinicians' need for justification. The findings reveal a certain level of inconsistency across works in terms of assessment criteria.

Turning to the aspect of anomaly detection, Fernando et al. present a survey of DL-based medical anomaly detection, whereas Yang et al. focus on DL methods for temporal data. Indeed, most temporal DL architectures for healthcare monitoring including autoencoders remain challenging to interpret. Sabic et al. discuss classical ML methods for heart-rate-based anomaly detection. Ray et al. and Schiff et al. investigate the use of statistical frameworks for outlier detection with applications to clinical decision support and medication error detection respectively. Another line of work explores the use of explainable ML for detecting data drifts as a means of identifying emerging risks, for example, Duckworth et al. study the use of explainable ML to identify health risks factors associated with emergency admission during the COVID-19 pandemic. There is also a growing body of work focusing on XAI approaches for IoMT and RPM security. Fahim and Sillitti provide a survey of general anomaly detection, analysis and prediction approaches on IoT, and some works have begun to address explainable intrusion detection approaches for medical IoT. In particular, lightweight ECG anomaly detection approaches that can operate with limited computational resources at the edge have been proposed, whereas signal processing or hybrid DL approaches have been used to secure IoMT device streams. Most works, however, focus on the development of detection models and have limited discussion of the need for explanations. Indeed, it has been recognized that in high-impact domains such as medical devices, detection alerts typically have to come with an explanation. Li et al. and works discussing explainable anomaly detection (XAD) address the need for interpretability in a broader sense and articulate the taxonomy of different explanation approaches, attention-based, perturbation-based and generative model-based ones. However, the majority of such works provide only high-level recommendations and are not specific to physiological time series.

A number of recent works address the application of post-hoc explainers to detect and explain different arrhythmias using ECG data. When applied to ECG-based arrhythmia detection SHAP, LIME, and Grad-CAM approaches reveal that gradient-based explanations correspond better to clinically relevant waveform features such as P-waves or QRS complex then do raw LIME scores, and that saliency-type explanations are generally preferable to counterfactuals by clinicians. The temporal nature of ECG data requires adaptations of standard LIME (which primarily assumes tabular input), resulting in proposals such as B-LIME and Sig-LIME for cardiac signal analysis. Another recent work by Zhang et al. proposes an XAI-based framework for detecting worsening depression and anxiety symptoms using consumer-level wearables: an LSTM autoencoder is trained on sleep, step counts and resting heart rate to detect clinically relevant anomalies in these variables.

III. FUNDAMENTALS OF XAI TECHNIQUES

XAI techniques applied to healthcare monitoring can be broadly grouped into model-agnostic, model-specific/visual, and counterfactual/rule-based approaches. Each offers a different trade-off between fidelity, computational cost, and the granularity of explanation delivered to the clinician.

A. Model-Agnostic Local Explanations: LIME and SHAP

Local Interpretable Model-agnostic Explanations (LIME), introduced by Ribeiro et al. [27], approximates the local decision boundary of any black-box classifier by fitting an interpretable surrogate model (e.g., linear regression) around a perturbed neighborhood of the instance being explained. SHAPley Additive explanations (SHAP), introduced by Lundberg and Lee [28], instead computes feature attributions using Shapley values from cooperative game theory, providing an attribution scheme with desirable consistency and local-accuracy properties. Both techniques have been applied extensively to ECG-based arrhythmia and atrial fibrillation classification [18]– [21]. However, standard LIME and SHAP implementations were originally designed for tabular, image, or text features and exhibit instability and limited local fidelity on temporally interdependent signal data; consequently, signal-adapted variants such as B-LIME [18], which incorporates heartbeat segmentation and bootstrapping, and Sig-LIME [19] have been proposed to improve reliability for one-dimensional biomedical signals.

B. Visual and Attention-Based Explanations: Grad-CAM and Attention Mechanisms

Gradient-weighted Class Activation Mapping (Grad-CAM), proposed by Selvaraju et al. [29], uses the gradients flowing into the final convolutional layer of a CNN to produce a coarse localization (saliency) map highlighting the regions of an input most responsible for a prediction. In ECG and time-series contexts, Grad-CAM has been adapted to highlight specific waveform segments (e.g., the QRS complex or ST segment) that drive a classification decision [16], [20]. Attention mechanisms, embedded directly within recurrent or transformer-based architectures, offer an alternative form of built-in interpretability by learning weights that indicate which time steps or channels the model attends to most strongly when producing a prediction; these have also been used for multimodal anomaly detection combining imaging and sequential clinical data [23].

C. Counterfactual and Rule-Based Explanations

Counterfactual explanations answer the question, 'what minimal change to the input would have altered the model's decision?' and can be intuitive for clinicians reasoning about intervention thresholds (e.g., the oxygen-saturation level at which an alert would not have fired). Rule-based and surrogate decision-tree explanations, meanwhile, extract human-readable if–then rules that approximate the black-box model globally or locally. Evidence from user studies of ECG explanation systems suggests that medical professionals generally prefer saliency-map-based explanations over counterfactual visualizations, citing clearer correspondence with established ECG-interpretation workflows [21], underscoring that explanation-format choice should be guided by clinical usability rather than technical convenience alone.

IV. ANOMALY AND EVENT DETECTION IN HEALTHCARE MONITORING

A. Data Modalities

Healthcare monitoring systems ingest heterogeneous data streams: quasi-periodic physiological waveforms (ECG, PPG), slowly varying vital signs (heart rate, blood pressure, SpO2, respiration rate), discrete behavioral/activity events from wearables (step count, sleep stage, fall events), and device- or network-level telemetry from IoMT infrastructure (access logs, packet traces, authentication events). Each modality imposes different constraints on anomaly detection and explanation: waveform data requires temporally localized explanations, vital-sign trends require drift- and threshold-aware explanations, and IoMT telemetry requires explanations grounded in network or protocol-level features rather than physiological ones [10]– [17].

B. Detection Approaches

Classical statistical approaches (e.g., control charts, Hidden Markov Models, spectral analysis) remain in use for lightweight, edge-deployable detection [14], [16], particularly where computational resources on wearable or embedded medical devices are limited. Supervised deep learning (CNNs, LSTMs, hybrid CNN-GRU architectures) dominates ECG and vital-sign anomaly classification where labeled event data is available [10], [18]. Unsupervised and self-supervised approaches, notably autoencoders and LSTM-autoencoders, are preferred when labeled anomalies are scarce, since these models learn a

representation of normal physiological or behavioral patterns and flag large reconstruction error as anomalous, as demonstrated in wearable-based mental-health monitoring [22]. Federated and hierarchical learning architectures have also been proposed to enable anomaly detection across distributed IoMT devices while preserving patient data privacy [25], [26].

V. PROPOSED CONCEPTUAL FRAMEWORK: XAI-INTEGRATED ANOMALY DETECTION PIPELINE

We propose a four-layer conceptual framework for embedding explainability into healthcare monitoring pipelines, synthesizing design patterns observed across the reviewed literature.

- Data Acquisition and Preprocessing Layer:

Ingests raw physiological and device-telemetry streams, applies noise filtering, artifact removal, segmentation (e.g., heartbeat or window segmentation), and class-imbalance mitigation such as SMOTE, ADASYN, or SMOTETomek, which have been shown to materially improve downstream ECG classification performance [15].

- Detection/Inference Layer:

A trained classifier or unsupervised detector (CNN, LSTM, autoencoder, ensemble) produces an anomaly score or event label for each input segment.

- Explanation Generation Layer:

One or more XAI techniques (SHAP, signal-adapted LIME, Grad-CAM, attention weights) are applied post-hoc or intrinsically to produce feature attributions, saliency maps, or attention weights aligned to the same temporal segment that triggered the detection.

- Explanation Aggregation and Presentation Layer:

Because point-wise attributions from SHAP/LIME are often too fine-grained for clinical interpretation, explanations are aggregated into clinically meaningful units (e.g., RR-interval or beat-level summaries) before being rendered to clinicians through a dashboard or alert interface, together with the raw signal segment and a confidence indicator [20].

This layered structure allows explanation techniques to be swapped or combined without redesigning the underlying detector, and it explicitly separates the concern of 'why was this flagged' from 'is this flag actionable', a distinction repeatedly identified as important by clinician user studies [21].

VI. APPLICATIONS AND CASE STUDIES

A. Cardiac Arrhythmia and Atrial Fibrillation Detection

The most mature application area combines CNN or ResNet-based classifiers with SHAP, LIME, and Grad-CAM to explain arrhythmia and atrial fibrillation predictions on datasets such as MIT-BIH, PTBDB, and the PhysioNet AFib dataset. Multi-level frameworks that combine LIME, SHAP, and Grad-CAM with RR-interval aggregation report strong precision/recall (e.g., 91.3% precision and 88.7% recall for AFib detection, and over 98% for normal rhythm classification in one study) while producing explanations aligned with established diagnostic criteria [20], [21]. Signal-specific LIME variants (B-LIME, Sig-LIME) further improve explanation stability over temporally dependent heartbeat data compared to unmodified LIME [18], [19].

B. IoMT Security and Device-Level Anomaly Detection

IoMT-connected monitoring devices are themselves subject to anomalous behavior arising from cyberattacks, device malfunction, or data poisoning. Anomaly-based intrusion detection systems combining CNN, LSTM, and attention mechanisms have been proposed for IoMT traffic classification [17], and lightweight spectral or hybrid HMM/SVM-based detectors have been designed for resource-constrained hospital and embedded medical devices [13], [14], [16]. Explanation in this setting typically takes the form of feature-importance rankings over network or protocol-level attributes rather than physiological waveforms, and remains comparatively underexplored relative to the cardiology literature.

C. Wearable-Based Behavioral and Mental Health Monitoring

Zhang et al. [22] developed an explainable anomaly detection framework using an LSTM autoencoder

trained on sleep duration, step count, and resting heart rate from consumer wearables across more than 2,000 participants, flagging anomalies corresponding to clinically significant increases in self-reported depression or anxiety scores. This case illustrates the extension of explainable event detection beyond acute physiological events toward longitudinal behavioral and mental-health trend monitoring, where explanations must communicate gradual multi-day deviations rather than a single waveform abnormality.

D. Elderly Care, Fall Detection, and Non-Wearable Monitoring

Edge-based IoT frameworks using non-wearable sensors have been proposed to detect behavioral anomalies in elderly patients, such as unusual inactivity or deviation from routine, to support caregivers and clinicians without requiring continuous wearable use. Combined with local, on-device processing, such systems can improve both privacy and latency, though explanation delivery in these contexts must be tailored to non-clinical caregivers rather than physicians alone.

VII. EVALUATION METRICS FOR EXPLAINABILITY AND DETECTION PERFORMANCE

Explainability methods can be evaluated from two perspectives: detection performance and explanation quality. With respect to the former, standard classification metrics such as accuracy, precision, recall/sensitivity, and F1 score, together with the area under the ROC curve for binary classification tasks, should be reported. With respect to the latter, the metrics vary depending on the explanation method. For model-agnostic approaches such as LIME or SHAP, fidelity can be measured in terms of how accurate the explanations are compared to the reference model, typically using the deletion or perturbation scores. Stability can be assessed as consistency of explanations across multiple runs or input samples, which tends to be a problem for default LIME implementation on temporal data. The clinical plausibility of explanations can also be evaluated, for instance, by comparing the explained features to established clinical criteria using domain knowledge or clinician feedback, and overall explanation quality can be ranked using user studies. It is important that an XAI-enabled application reports both detection and explanation performance.

VIII. CHALLENGES AND OPEN RESEARCH ISSUES

- Fidelity–interpretability trade-off: Simpler, more interpretable models (e.g., decision trees, linear models) often underperform deep architectures on high-dimensional physiological data, while highly accurate models require post-hoc explanation methods that only approximate, rather than reveal, the true decision process.
- Explanation instability on temporal data: Perturbation-based methods such as LIME were designed for tabular, text, and image data and exhibit instability when applied naively to temporally interdependent signals, motivating signal-specific adaptations such as B-LIME and Sig-LIME [18], [19].
- Granularity mismatch: Point-wise feature attributions are frequently too fine-grained for clinical decision-making, which is typically based on morphological or interval-level patterns rather than individual sample points, necessitating aggregation strategies [20].
- Lack of standardized evaluation protocols: There is no consensus benchmark or metric suite for jointly assessing detection accuracy and explanation quality across healthcare monitoring tasks, which complicates cross-study comparison [3].
- Privacy, security, and regulatory compliance: Explanation outputs themselves may leak sensitive patient information, and regulatory frameworks such as HIPAA, GDPR, and the EU AI Act impose additional constraints on how automated healthcare decisions must be documented and justified [2].
- Underexplored security-oriented explainability: While cardiology applications of XAI are relatively mature, explanation techniques for IoMT device-level and network-level anomaly detection remain comparatively underdeveloped [11]– [17].

IX. FUTURE RESEARCH DIRECTIONS

- Development of intrinsically interpretable temporal architectures (e.g., attention-based or prototype-based sequence models) that avoid the instability of post-hoc perturbation methods on physiological

signals.

- Standardized, clinician-validated benchmarks that jointly score detection accuracy, explanation fidelity, and explanation stability across shared healthcare monitoring datasets.
- Integration of large language models (LLMs) with anomaly detectors to translate technical feature attributions into natural-language clinical narratives, extending early work such as HuntGPT in the cybersecurity domain [23] to clinical monitoring contexts.
- Federated and privacy-preserving explainability, enabling multi-institutional model training and explanation generation without centralizing sensitive patient data [25], [26].
- Human-centered evaluation studies involving practicing clinicians and caregivers to determine which explanation formats (saliency maps, counterfactuals, natural-language summaries) most effectively support timely and correct clinical decisions across diverse monitoring scenarios.
- Explainable anomaly detection for IoMT device security, extending mature cardiology-focused XAI techniques to protect the monitoring infrastructure itself against tampering, spoofing, and adversarial data poisoning.

X. CONCLUSION

Explainability is a critical requirement for the clinical use of automated anomaly and event detection in healthcare, particularly in the context of bedside and wearable monitoring. This paper provides an overview of the state-of-the-art in this field by reviewing the main XAI methods (LIME, SHAP, Grad-CAM, attention mechanisms, counterfactuals) and discussing their applications. We detail major opportunities and challenges for applying these methods in the healthcare monitoring context and articulate a conceptual framework for developing such applications. While the works reviewed in this paper demonstrate that adapted versions of tabular or image explanation methods can yield effective results for ECG-based arrhythmia detection, the application of XAI methods to other physiologic signals and IoMT device-level security requires further research. Particular challenges include the need for temporal adaptations of model-agnostic methods and development of clinically useful explanation formats.

REFERENCES

[1] Adadi and M. Berrada, “Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[2] I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel, and B. Ogbuokiri, “A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges,” Informatics in Medicine Unlocked, vol. 51, p. 101587, 2024.

[3] J. Allgaier, L. Mulansky, R. C. Draelos, and R. Pryss, “How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare,” Artificial Intelligence in Medicine, vol. 143, p. 102616, 2023.

[4] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, “Deep learning for medical anomaly detection—A survey,” ACM Computing Surveys, vol. 54, no. 7, pp. 1–37, 2021.

[5] X. Yang et al., “Deep learning technologies for time series anomaly detection in healthcare: A review,” IEEE Access, vol. 11, pp. 117788–117799, 2023.

[6] E. Sabic, D. Berbakov, and A. Vasic, “Healthcare and anomaly detection: Using machine learning to predict anomalies in heart rate data,” AI & Society, vol. 36, no. 1, pp. 149–158, 2021.

[7] S. Ray, S. R. Kannan, and S. Sengupta, “Using statistical anomaly detection models to find clinical decision support malfunctions,” Journal of the American Medical Informatics Association, vol. 25, no. 7, pp. 862–871, 2018.

[8] G. D. Schiff et al., “Screening for medication errors using an outlier detection system,” Journal of the American Medical Informatics Association, vol. 24, no. 2, pp. 281–287, 2017.

[9] C. Duckworth et al., “Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19,” Scientific Reports, vol. 11, no. 1, p. 23017, 2021.

[10] G. Sivapalan, K. K. Nundy, S. Dev, B. Cardiff, and D. John, “ANNet: A lightweight neural network for ECG anomaly detection in IoT edge sensors,” IEEE Transactions on Biomedical

Circuits and Systems, vol. 16, no. 1, pp. 24–35, 2022.

[11] X. Wu, Y. Yang, M. Bilal, L. Qi, and X. Xu, "6G-enabled anomaly detection for metaverse healthcare analytics in Internet of Things," IEEE Journal of Biomedical and Health Informatics, 2023, doi: 10.1109/JBHI.2023.3298092.

[12] M. Fahim and A. Sillitti, "Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review," IEEE Access, vol. 7, pp. 81664–81681, 2019.

[13] A. M. Said, A. Yahyaoui, and T. Abdellatif, "Efficient anomaly detection for smart hospital IoT systems," Sensors, vol. 21, no. 4, pp. 1–24, 2021.

[14] N. Sehatbakhsh, M. Alam, A. Nazari, A. Zajic, and M. Prvulovic, "Syndrome: Spectral analysis for anomaly detection on medical IoT and embedded devices," in Proc. 2018 IEEE Int. Symp. Hardware Oriented Security and Trust (HOST), 2018, pp. 1–8.

[15] M. Zhang, A. Raghunathan, and N. K. Jha, "MedMon: Securing medical devices through wireless monitoring and anomaly detection," IEEE Transactions on Biomedical Circuits and Systems, vol. 7, no. 6, pp. 871–881, 2013.

[16] M. Yamauchi, Y. Ohsita, M. Murata, K. Ueda, and Y. Kato, "Anomaly detection in smart home operation from user behaviors and home conditions," IEEE Transactions on Consumer Electronics, vol. 66, no. 2, pp. 183–192, 2020.

[17] M. M. Khan and M. Alkhathami, "Anomaly detection in IoT-based healthcare: Machine learning for enhanced security," Scientific Reports, 2024, doi: 10.1038/s41598-024-56126-x.

[18] T. A. A. Abdullah, M. S. M. Zahid, W. Ali, and S. U. Hassan, "B-LIME: An improvement of LIME for interpretable deep learning classification of cardiac arrhythmia from ECG signals," Processes, vol. 11, no. 2, p. 595, 2023.

[19] T. A. A. Abdullah, M. S. M. Zahid, A. F. Turki, W. Ali, A. A. Jiman, M. J. Abdulaal, N. M. Sobahi, and E. T. Attar, "Sig-LIME: A signal-based enhancement of LIME explanation technique," IEEE Access, vol. 12, pp. 52641–52658, 2024.

[20] J. Luo, T. Amirsajjad, P. Noffke, and R. Sparapani, "Multi-level explainable AI for ECG-based atrial fibrillation detection: Exploring LIME, SHAP, and Grad-CAM for clinical interpretation," in Proc. 10th ACM/IEEE Symposium on Edge Computing, 2025, pp. 52–59.

[21] "Explainable AI (XAI) for arrhythmia detection from electrocardiograms," arXiv preprint arXiv:2508.17294, 2025.

[22] Y. Zhang et al., "An explainable anomaly detection framework for monitoring depression and anxiety using consumer wearable devices," arXiv preprint arXiv:2505.03039, 2025.

[23] M. M. Ali and V. Kostakos, "HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models," arXiv preprint arXiv:2309.16021, 2023.

[24] Z. Li et al., "A survey on explainable anomaly detection," ACM Transactions on Knowledge Discovery from Data, vol. 18, no. 1, 2023.

[25] J. Cawthra et al., "Securing telehealth remote patient monitoring ecosystem," NIST Special Publication 1800-30B, National Institute of Standards and Technology, 2022.

[26] "Hierarchical federated learning-based anomaly detection using digital twins for smart healthcare," arXiv preprint arXiv:2111.12241, 2021.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.