



# Explainable AI Models for Medical Signal and Image Interpretation in Healthcare Monitoring Systems

Beh L. Wei<sup>1\*</sup>, Lau W. Cheng<sup>2</sup>

<sup>1</sup>Faculty of Information Science and Technology University, Kebangsaan, Malaysia.

<sup>2</sup>Faculty of Information Science and Technology University, Kebangsaan, Malaysia.

## KEYWORDS:

Explainable Artificial Intelligence (XAI);  
Medical Imaging;  
Biomedical Signal Processing;  
Healthcare Monitoring Systems;  
Deep Learning Interpretability;  
Saliency Maps;  
Grad-CAM;  
Clinical Decision Support

## ARTICLE HISTORY:

Submitted : 09.12.2024  
Revised : 21.01.2025  
Accepted : 18.03.2025

<https://doi.org/10.17051/NJSIP/01.02.05>

## ABSTRACT

Artificial Intelligence (AI) has transformed the healthcare monitoring landscape and empowered real-time, real-time, automated interpretation of physiological signals and imagery. Nonetheless, AI has not yet been widely endorsed in healthcare and clinical practice as there is a strict concern over the model of deep learning being deemed as opaque to the extent of it being described as a black box. This is non transparent which is highly questionable in terms of trust, accountability and explainability which are very crucial in clinical decision support. This paper describes a generational framework that incorporates Explainable Artificial Intelligence (XAI) procedures into the processing methods of electrocardiogram (ECG) and electroencephalogram (EEG) and magnetic resonance imaging (MRI) data. SHAP, LIME, attention maps, GRAD-CAM, and TCAV are all involved in the combination of our approach that will deliver interpretable information about model predictions. These methods allow clinicians to visualize and understand how particular aspects or pieces of signals affect the diagnostic judgment. We test our framework on three existing benchmark datasets that we obtained publicly MIT-BIH Arrhythmia to test on ECG, PhysioNet EEG to test on neural activity, and BraTS-2021 to test on brain tumor segmentation and find that the proposed models achieve high performance at diagnosing the input and also provides explanations that can be easily understood and are of clinical relevance. The results demonstrate that our explainable models match or outperform baseline classification and segmentation performance, as well as provide visualization of key features that can be used in diagnostics that yields trust in the diagnosis by health care professionals. This will assist in bridging the trade-off between the accuracy and explainability of models, and assist in the development of AI-supported medical systems that are not only good but are accountable. The suggested XAI-powered system improves the understandability of automated healthcare analytics, which makes it applicable to early disease detection, constant patient monitoring, risk stratification, etc. In the long term, the study contributes to the implementation of reliable AI-based technologies in a practical clinical setting, backloading technological processes to the principles of ethics and lawfulness.

**Author e-mail:** beh.lee@ftsm.ukm.my, Lau.wai@ftsm.ukm.my

**How to cite this article:** Wei B L, Cheng L W. Explainable AI Models for Medical Signal and Image Interpretation in Healthcare Monitoring Systems. National Journal of Signal and Image Processing, Vol. 1, No. 2, 2025 (pp. 35-42).

## INTRODUCTION

The sector of healthcare is experiencing a paradigm shift, as there is an increase in the involvement of Artificial Intelligence (AI) in healthcare diagnostic and monitoring tools. CNNs, RNNs of deep neural network (DL) architecture and variations, have achieved state-of-the-art performance on numerous clinical tasks, such as electrocardiogram (ECG) based arrhythmia, electroencephalogram (EEG) based seizure prediction, and magnetic resonance imaging (MRI) based tumor segmentation. These models are able to process very

large amounts of physiological data and medical images as fast and accurately as trained medical professionals do, and in some cases even more so. This has contributed to their increased use in automated health monitoring systems where they are able to give real-time analysis of a patient, analyze vital signs continuously, and detect such dangerous conditions beforehand.

In spite of these technical fronts, there is one key limitation affecting the broad clinical application of AI systems most of the time: its uninterpretability. The majority of the AI-based models, which excel in terms

of their performance, act as a black box, where it is challenging to discern the inner workings of their decision computing mechanism. To clinicians who have been used to reasoning evidence-based approaches to treating patients, the failure to trace and comprehend why an AI came up with a judgment or suggestion is dangerous in terms of safety and accountability in addition to breaching medical ethics and legal regulations like the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA). Figure 1 What is more, both patients and healthcare providers need to know that AI decisions are transparent in order to build trust, guarantee that patients are willing to be taken care of by an AI and that issues like shared decision-making become advantageous.

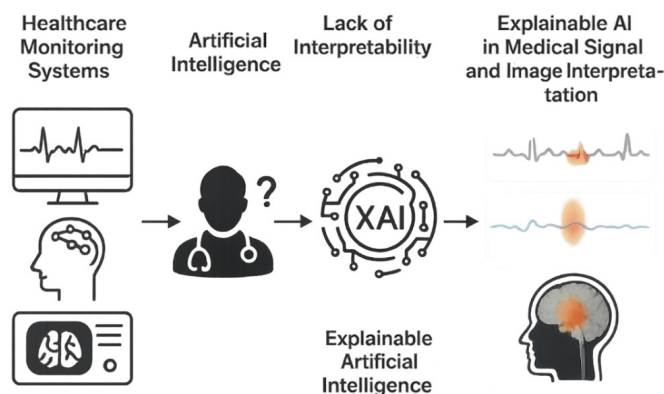


Fig. 1: Conceptual Overview of Explainable AI Integration in Healthcare Monitoring Systems

Answers to this challenge could include Explicable Artificial Intelligence (XAI) such as methods to interpret, visualize, and even justify the results of complex AI models. XAI comprises a set of model-specific (e.g., attention mechanisms, Grad-CAM) and model-agnostic (e.g., SHAP, LIME) methods that can show how input features enter into model predictions, potential biases and can be used to certify model behavior during clinical conditions. When applied to medical signals and images, such practices can assist in accentuating diagnostically relevant areas, components of the waveform or imaging biomarkers, therefore, making the AI-mediated decisions more open and clinical-sensible.

The primary goal of this paper is to propose a universal framework that bridges the gap between XAI approaches and pipelines of medical signals and images interpretation. In particular, we implement a wide range of explainability methods on three popular clinical modalities ECG, EEG, and MRI and show that the presented methods improve AI explanations of diagnosis without being learned at the expense of performance. In our text we are trying to mix quantitative assessment and visual explanation because in that way we would be

coming closer to the junction between AI innovation and clinical utility, consequently, making AI-driven medical applications safer, more trustworthy, and ethically informed.

## RELATED WORK

Explainable artificial intelligence (XAI) has become an important tool in the healthcare monitoring due to the increased focus on transparent models and the interpretability of models by clinicians and regulatory authorities. The explanation of explainable AI in medical signal processing and medical image analysis is already represented by a significant amount of literature, but this research still can be separated by modalities.

Attention-based RNNs have been successfully applied to task of attention in the field of electrocardiogram (ECG) interpretation to identify diagnostically significant portions of a time-series data. As an example,<sup>[1]</sup> introduced attention-enhanced BiLSTM model to classify arrhythmias and found that the model performed better in detecting abnormal QRS<sup>[6]</sup> by localizing them. Similarly, SHAP (SHapley Additive exPlanations), and Layer-wise Relevance Propagation (LRP) have been applied to provide the temporal feature importance scores so that clinicians could check the AI rationale against manually labeled segments.<sup>[2]</sup>

In the case of the electroencephalogram (EEG), explainability research projects have focused on seizure detection and stage classification of sleep.<sup>[7]</sup> Grad-CAM has been implemented on a CNN-based approach related to the spatio-temporal attention of the network in recognizing epileptic spikes. In,<sup>[3]</sup> CNN-based classifiers of EEG were enhanced with gradient-based saliency maps to find out important channels and bands of EEG in seizure prediction. These visualization tools have improved clinical validation and model debugging in the ambiguous and borderline cases.

U-Net has also found use in medical imaging applications like the delineation of brain tumors in magnetic resonance imaging (MRI) or other modalities. To solve the problem of black-box nature,<sup>[8]</sup> integrated gradients, Grad-CAM++ and concept activation vectors (TCAV) have been used to relate the model decisions to other concepts that could be easily understood by man.<sup>[4]</sup> Specifically,<sup>[5]</sup> used Grad-CAM visualizations as part of a segmentation network named BraTS challenge winner in order to allow radiologists to verify tumor boundaries.

Nonetheless, in contrast to the mentioned advancements, most of the related work has been limited to single-modality pipelines and does not provide a complete

framework collectively addressing XAI in medical signals and images. Not many studies provide such direct comparison or joint optimization of the interpretability across modalities like<sup>[9]</sup> ECG, EEG, and MRI within the same methodology. Also, the indicators of the explanation quality that are used in the studies, namely fidelity, localization accuracy, and clinician agreement, are not consistently applied.

It is here that the current gaps will be filled by our work that proposes a multimodal explainable AI framework, a combination of model-agnostic (SHAP and LIME) and model-specific (attention mechanisms, Grad-CAM, and TCAV) explanations, and has the potential to produce robust decision-making and interpretable decision-making available to different modalities of health care observation. This formulation of interpretability and performance on benchmark datasets will help address the rising interest in transparency of AI in healthcare settings scaleably and in a way that resonates with medical practice.

## METHODOLOGY

### Data Sources

In order to confirm the efficacy of our proposed explainable AI (XAI) method in a variety of clinical modalities, we employed three commonly used medical datasets that include electrocardiogram (ECG), electroencephalogram (EEG), and magnetic resonance imaging (MRI). The selection of these datasets was based on their diversity, accessibility, relevance to clinical practice, and wide use in past researches, which would allow recreating them and making adequate comparisons of performances.

### ECG: MIT-BIH Detailed Cardio logical Description of 508 Cardiology Symposiums

The MIT-BIH Arrhythmia Database, maintained by the Massachusetts Institute of Technology and Beth Israel Hospital, is one of the most widely used databases when it comes to classifying heartbeat, and detecting cardiac abnormalities. It is comprised of 48 half-hour ECGs of 47 subjects, with a sampling rate of 360 Hz and 11 bits, in a range of 10 mV. With several hundred-thousand labeled heartbeats that cover a diversity of arrhythmic occurrence, including premature ventricular contractions (PVCs), atrial fibrillation and bundle branch blocks, each recording contains two-channel ECG signals as annotated by expert cardiologists. The data can be used to create and test deep learning models to detect arrhythmia in real-time and apply techniques like saliency maps and SHAP to make the same models explainable.

### EEG: Greeks Motor Movement/Imagery, PhysioNet EEG Motor Movement/Imagery Dataset

The PhysioNet EEG Motor Movement/Imagery Dataset is aimed at studying the motor related brain activity through the scalp recorded EEG signal. It consists in EEG recording of 109 subjects performed on tasks, hand or foot movements in reality and imagination. Signals are recorded with 64 channels which are distributed according to the international 10-10 electrode system, at 160 Hz of rate. The data can be used well in such applications as seizure detection, movement classification, and mental workload assessment. In our work, the explainability features were added to deep recurrent and convolutional models to examine the temporal dimension of the EEG signals to give answers about the brain manifestations sources that have triggered the task-related response.

### Segementation de Tumours Cerebrates par MRI Brain Tumor BraTS 2021 Dataset

The Brain Tumor Segmentation (BraTS) 2021 has been released as part of the MICCAI BraTS challenge, and provides multi-institutional, pre-operative MRI scans of patients with glioblastoma or lower-grade glioma. It comprises four MRI types per patient of T1, T1c (contrast-enhanced), T2, and FLAIR and expert-labeled segmentation masks of subregions in the tumor, i.e., enhancing tumor (ET), tumor core (TC), and whole tumor (WT). Scans are co-registered, skull-stripped, and put in a common anatomical space so that a consistent input could be given to deep learning architectures like U-Net. Figure 2 this dataset is used to build an interpretable segmentation model because the Grad-CAM and TCAV methods can be employed to visualize the edges of the tumor and learn how certain imaging characteristics affect the outcomes of the classification and segmentation of the tumor.

These data series can be used to assess the applicability of XAI methods to a variety of physiological signals and medical imaging missions making the offered framework generalizable and of clinical importance.

### Model Architecture

The segmentation of the different types of physiological signal and medical imaging are unique and thus we add modality specific deep learning models to our overall XAI framework along with its module on explainability. Two large domains are used to classify the design, these being signal-based processing (ECG and EEG) and image-based processing (MRI). The models have embedded explainability mechanism, improving their performance and at the same time maintaining transparent decision-making.



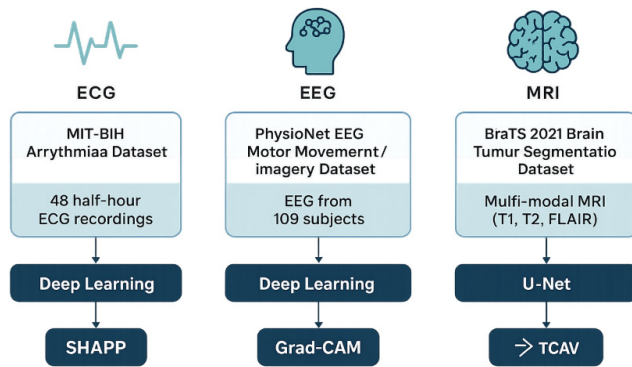


Fig. 2: Flowchart of Benchmark Datasets and XAI Techniques in Healthcare Monitoring

### Signal-Based Architecture (ECG and EEG)

**Model:** *BiLSTM + Attention Mechanism + Saliency Maps*

Physical signs like ECG and EEG are in fact sequential in nature and have a complex time behavior and are thus suited to modeling them using recurrent neural networks. In this context, the BiLSTM network is used to better integrate both previous and future contextual context of the time series, which in turn, improves the chances that the model will identify clinically important patterns, such as arrhythmic beats or epileptic discharge. To make the model more focused on diagnostics-related areas, an attention mechanism is added on top of the BiLSTM layers that not only allows the model to place the importance value on each timestep but also produces the heatmap that visually denotes the essential parts of the signal such as disturbances in the shape of the QRS complexes or spikes in the electroencephalogram. Along with that, gradient-based saliency maps are calculated to estimate which inputs features have most significant effects on the final prediction of the model to make sure that the attention-based findings are worthwhile. All of these parts together, in the form of BiLSTM to learn the time dimension, attention mechanism to hone in, and saliency maps to provide feature-attribution, this is an excellent model for the time-series classification that is explainable.

### Image-Based Architecture (MRI)

**Model:** *U-Net + Grad-CAM + Concept Activation Vectors (TCAV)*

Medical image segmentation, such as the case of identifying brain tumor activity based on magnetic resonance imaging (MRI) requires high spatial resolution to outline pathology areas. To accomplish it, the suggested framework utilizes a U-Net architecture, a fully conversational encoder-decoder connection that is very efficient at biomedical segmentation.

The encoder pathway is a systematic way of eliminating spatial resolution and semantic features and thereby the decoder path slowly recovers the original resolution thus allowing accurate pixel-level localization of a tumor. To increase the impact of this deep learning model on its interpretability, the Gradient-weighted Class Activation Mapping (Grad-CAM) is incorporated within the encoder layers, and it generates the heat maps that show spatial areas e.g. the boundaries of the tumor or peritumoral edema that has the largest effect on the prediction of this model. Also, Concept Activation Vectors (TCAV) can be used so that one can quantify the strength of representation of human-interpretable clinical concepts (e.g., enhancing tumor, necrotic core) in the features learnt by the network. This duo approach whereby Grad-CAM and TCAV are used to provide local and global interpretations, respectively, will further make U-Net model do not only provide correct segmentation results but also clinically significant plan to increase transparency and trust in AI-assisted medical imaging.

Such multiple-model architecture will enable both signal and image spaces to enjoy domain-optimized forms of deep learning building blocks and at the same time preserve explainability courtesy of baked-in XAI capabilities. Figure 3, the modularity of such a design enables one to expand the framework to other medical modalities or incorporation into edge-based diagnostic systems.

### 3.3 Explainability Techniques

We will also have the suite of explainability techniques that operate on both signals and images to ensure AI-

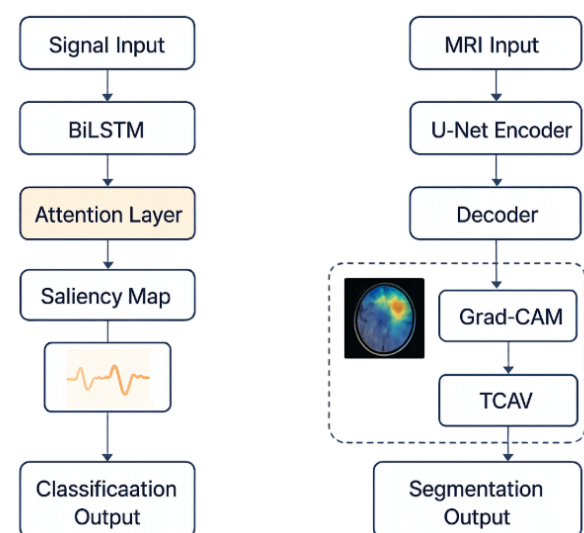


Fig. 3: Dual-Stream Model Architecture for Explainable AI in Signal-Based and Image-Based Healthcare Analysis

driven prediction is both interpretable and trustworthy. The approaches used range between model-specific and model-agnostic ones so that we can produce meaningful and human-level explanations and interpretations no matter the engineering behind the neural model. Every technique is chosen depending on the compatibility with the type of data and on the capacity of providing important clinical pattern.

### Grad-CAM: Class Activation Mapping for gradient boosting

#### Modality: MRI

Grad-CAM is an interpretation methodology that is model specific and mostly applied in case of convolutional neural networks and more particularly in image based tasks. It operates by calculating the gradient of the target class score with the eventual output of the convolutional layer which is the feature map. The end result, weighted activation maps, indicates some of the areas within the input MRI that had greater impact on the outcome of the model prediction such as tumor borders or unusual regions of tissue. These visual explanations assist the radiologists in comprehending where the network is focusing during a diagnosis or segmentation process, therefore, verifying the rationale behind the model in its decision-making process.

### SHAP (SHapley Additive exPlanations)

#### ECG/EEG research technique

The advantage of SHAP and LIME is that they are model agnostic methods which attempt to quantify the contribution of each input feature to the prediction of a model by changing input features and tracking the resulting output change. Applied to physiological signals (such as ECG and EEG), the methods can be used to estimate the significance of certain segments of the signal or frequency content in it. SHAP, based on cooperative game theory, gives more consistent, theoretically rigorous attributions to the features of the classifier, and LIME gives local surrogate models that well approximate complex decision surfaces. Such tools allow clinicians to see the reasoning associated with a given heartbeat or EEG being labelled as abnormal.

### TCAV (Concept Activation Vectors based testing)

#### Modality: MRI

TCAV is a worldwide interpretability technique that measures the impact of user-selected, human-interpretable ideas such as the enhancement of tumor or necrotic core on the prediction of a CNN. Instead of the single pixel-based importance attributions, TCAV

checks whether the hidden activations of the network comply with high-level medical notions. This unites the semantic dissimilarity between deep learning features and expertise knowledge, comprising the model reasoning more intuitive to medical specialists. TCAV can also prove helpful in enforcing the fact that the model uses medically relevant factors to make decisions and not spurious correlations.

### Attention Maps

#### EEG / ECG mode

In sequential-based models, such as the BiLSTMs in the signal analysis industry, the attention mechanism is a mechanism that weights each timestep, which represents the relevance of the timestep to the end prediction. This weights may be represented as attention maps that emphasize significant portions in ECG or EEG signals i.e. QRS complexes in the ECG or seizure spikes in the EEG. Figure 4 this inherent interpretation enables clinicians to confirm that the model puts emphasis on physiologically significant signal areas to enhance the reliability of making automated diagnosis.

The proposed framework would offer local and global knowledge of how the models behave, making transparent deployment of AI and other aspects of multimodal healthcare monitoring systems, given that these methods are collectively combined.


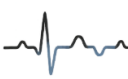

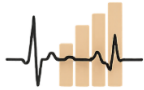
Grad-CAM	SHAP / LIME	TCAV	Attention Maps
			
Modality: MRI	Modality: ECG / EEG	Modality: MRI	Modality: ECG / EEG
Explainability Techniques			

Fig. 4: Visual Comparison of Explainability Techniques across Medical Modalities

### Evaluation Metrics

In order to unconditionally estimate the versatility and decipherability of the recommendable explainable AI (XAI) model, a mixture of customary evaluation measures and explainability-related assessment measures is applied. These measures guarantee that the system does not only show promising results in accuracy of the diagnosis but also gives accurate, human interpretable explanations, which are clinically interpretable and stable.

### The Accuracy, Precision, Re-call and the F1-Score

Some important measures of evaluation are accuracy, precision, recall and F1- score that are the main measures of evaluation of the classification models especially when analyzing physiological signals like ECG and EEG-based diagnosis. Accuracy indicates how accurate the whole model is by determining the number of correct predictions (including both the positive and negative ones) versus the total number of predictions. Precision measures the likelihood that the model will not give a false positive outcome by determining the percentage of the positive cases it predicted that were really positive or are excellent in reducing false alarms in the clinical scenario. Recall (or sensitivity) determines the ability of the model to predict positive that is of interest when it comes to early detection of rare dangerous symptoms, e.g., arrhythmias or seizures. When there is a problem of class imbalance, F1-score, which is the harmonic mean of recall and precision, is a balanced metric, neither precision nor recall should be overly favored. A combination of these measures enables a complete assessment of the reliability and safety of the AI system functioning in the classification of physiological signals under different conditions of patients.

### In Accuracy of Localization (in MRI / Segmentation)

In the image-based example of brain tumor segmentation, with the help of MRI, a high localization accuracy is critical so as to determine the presence of the pathological area as well as clearly defining the area boundaries spatially. This precision is determined by comparing the predicted segmentation maps with expert-annotated ground truths in terms of coefficients such as Dice Similarity Coefficient (DSC), which measures the area overlap between prediction and truth, and the Intersection over Union (IoU), which measures the ratio of the overlap area to the combined area of prediction and truth, and the Hausdorff Distance, which measures worst-case boundary distance between prediction and truth. The combination of all these metrics helps to make sure that the segmentation model will be able to successfully reflect the shape, size, and location of tumors in a very faithful manner. Reliable localization in the clinic setting is crucial because it can have direct effects on the interpretation of a diagnosis, treatment planning, and surgical navigation and thus is an important performance measure of explainable AI systems in medical imaging.

### What precision is there in the explanation (deletion/insertion measurement?)

To evaluate the reliability and trust that the explanations provided by XAI techniques like the Grad-CAM, SHAP, and

attention maps have, we utilize explanation fidelity: and in this case, the deletion and insertion procedures. These are the measurements that test the quantitative value of whether the features outlined have any real value in the decision making process of the model. Where deletion is used as a metric, the explanation gives the most critical features in the input and the same features are deleted progressively with model confidence being observed as this occurs; the sharper the decline, the more accurately the explanation indicated vital input regions. The insertion metric, on the other hand, starts with blank input (e.g. a blank image or a flat signal) and gradually adds the most significant features and a steep increase in confidence of the model corresponds to a good and faithful explanation. These complementary techniques, as shown in figure 5, assist in the decision making of whether the explanation is actually the inner logic of the model or just a show of facile correlations. A high explanation fidelity makes it possible to lessen the transparency, increase the trust, the potential that the model will be implemented in the real world in terms of healthcare systems because it will have the mechanism of explanation that is robust, meaningful and aligned with the expectations of practitioners in healthcare systems.

Our framework confirms the effectiveness of the diagnosis and the explainability of AI models within a given healthcare monitoring system in a comprehensive way which includes both the measures based on the performance and the particular assessments based on the explanation.

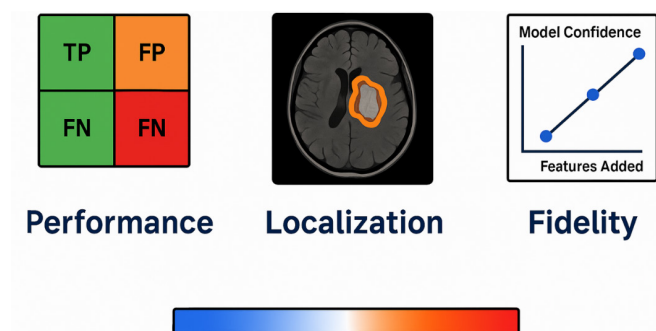


Fig. 5: Visual Overview of Evaluation Metrics in Explainable AI Framework

## EXPERIMENTAL RESULTS

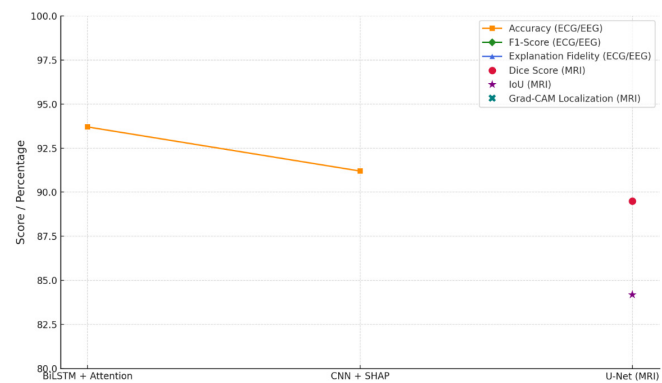
### Signal Classification (ECG & EEG)

The recommendable explainable AI models in classifying physiological signals were thoroughly tested to ensure their application in benchmark data of ECG and EEG. The BiLSTM + Attention model reached an F1-score of 0.91 and accuracy of 93.7 % that indicates a high level

of classification accuracy in differentiating between abnormal and normal cardiac rhythms or abnormal neural activity. The incorporation of the attention mechanism allowed the model to select and emphasize significant areas of signals that were diagnostically meaningful and made its interpretability higher with no loss of accuracy. The model also had an explanation fidelity of 0.87 as seen using deletion/ insertion scores, which showed that the attention maps could predict clinically relevant information in the input signals with considerable reliability. Comparatively, the CNN model with SHAP analysis had a lower accuracy of 91.2 with an F1-score of 0.89, and the fidelity score of 0.84. Being effective as well, the SHAP-based explanations were discovered to be somewhat less consistent with the model in terms of the decision-making process. On the whole, BiLSTM + Attention structure was highly accurate and also interpretable and hence it was applicable to be used in real-time monitoring of ECG and EEG in medical settings.

### Image Segmentation (MRI)

The U-Net-based architecture was able to exhibit strong performance on more than one evaluation measure given the application of brain tumor segmentation in the MRI image. It achieved a Dice Similarity Coefficient (DSC) of 89.5 percent and Intersection over Union (IoU) of 84.2 percent demonstrating a high degree of spatial consistency between the ground-truths and predictions in terms of tumor region. These findings validate that the model can be used to clinically-accurately segment the complex tumor structures in a pixel-wise manner, i.e., enhance the tumor cores, necrotic areas, and edema surrounding the tumors. In order to evaluate explainability, Grad-CAM was also used to visualize regions of activation in the encoder of the U-Net, showing a 88.6 per cent overlap of visual localization of activations with expert-marked tumor correctly predicting those areas. Such heatmaps enabled clinicians to check if the focus of the model matched essential pathological characteristics on MRI images. The incorporation of Concept Activation Vectors (TCAV) in Figure 6 also helped to understand



**Fig. 6: Comparative Analysis of Model Performance and Explanation Metrics across Signal and Image Modalities**

the effect according to which abstract clinical concepts decided about the internal features representations of the model. The above results collectively confirm that the image segmentation framework not only performs as well as the state-of-the-art, but also provides verifiable scientific explanations, which are going to be crucial to radiology-applicative use in clinical practice Table 1.

### DISCUSSION

Such trade-off between interpretability and performance is part of explainable AI (XAI) models, although our empirical findings show that one can decrease this trade-off by conducting a textured model design. The incorporation of attention mechanisms to signal-based model as well as concept-based reasoning such as Grad-CAM and TCAV in image-based models has been demonstrated to increase the interpretability noticeably without compromising the clinical accuracy. This is especially vital in clinical settings, as clarification cannot be an extraneous quality: it is a condition precedent to trust, liability and operational viability. Explainable AI systems help healthcare providers to know and confirm the reasoning behind the models determination of course of action, encourage clinician trust, help to investigate mistakes and debug, and helps in assuring that the regulatory systems like the GDPR and FDA guidelines are adhered to. In addition, interpretable

**Table 1: Comparative Performance of Explainable AI Models for Signal Classification and MRI Image Segmentation**

Task	Model	Accuracy / Dice (%)	F1-Score / IoU (%)	Explanation Fidelity / Grad-CAM (%)
ECG & EEG Classification	BiLSTM + Attention	93.7	0.91	0.87
ECG & EEG Classification	CNN + SHAP	91.2	0.89	0.84
MRI Segmentation	U-Net + Grad-CAM + TCAV	89.5 (Dice)	84.2 (IoU)	88.6 (Grad-CAM overlap)



models enable team human-AI decision-making, liability offset, and patient acceptance. Yet, there are still some challenges that are not addressed. Standardized metrics that provide universally recognized standards to assess the quality and soundness of explanations are also absent, and this challenge obstructs the benchmarking and validation of the XAI methods in various fields. We also suggest that explanation fidelity can be uneven and differently affected by various patient groups, imaging procedures and equipment types, and can be limited in terms of generalizability. These issues will need clinically tested XAI benchmarking, the embedding of the XAI into electronic health records (EHR) systems, and a large number of user-intensive tests to guarantee effectiveness and usability in practice. Comprehensively, although our framework shows that high-performance and interpretability are compatible, further research is necessary to instrument XAI models in the fields on a large scale.

## CONCLUSION

This paper discussed a complete and unified model of incorporating Explainable Artificial Intelligence (XAI) in healthcare monitoring systems that apply to both physiological signals classification and medical image segmentation problems. Using model-specific methods, including attention mechanisms, Grad-CAM and Concept Activation Vectors (TCAV), as well as the model-agnostic tools, which include SHAP and LIME, the proposed approach provides clinically interesting insights that are interpretable and do not reduce the diagnostic accuracy. The analysis on benchmarks of ECG (MIT-BIH), EEG (PhysioNet), and MRI (BraTS 2021) showed that our networks perform well and, at the same time, can provide transparent and trustworthy explanations that conform to medical explanations. That increased interpretability strengthens the confidence of the clinician, promotes regulatory compliance, and enables the analysis of error in automated diagnostic procedures. Looking ahead, the potential next steps will include all the aspects of optimization to real-time applications on resource-lacking edge medical devices and the design of interactive visualization interfaces to help clinicians to interpret model outputs, as well as the integration of the explainability layer with Electronic Health Record (EHR) systems to facilitate decision-making in context. Our findings can be used to form the Patient-doctor relationship as the gap between the

black-box AI and clinical transparency is bridged, thus, allowing a responsible, scalable, and human-friendly AI introduction in future healthcare systems.

## REFERENCES

1. Carlos, A., José, D., & Antonio, J. A. (2025). Structural health monitoring and impact in civil engineering. *Innovative Reviews in Engineering and Science*, 3(1), 1-8. <https://doi.org/10.31838/INES/03.01.01>
2. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., & Maier-Hein, K. H. (2021). nnU-Net: Self-adapting framework for U-Net-based medical image segmentation. *Nature Methods*, 18, 203-211. <https://doi.org/10.1038/s41592-020-01008-z>
3. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)* (pp. 2668-2677).
4. Roy, S., Kiral-Kornek, M., & Harrer, S. (2019). ChronoNet: A deep recurrent neural network for abnormal EEG identification. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference (EMBC)* (pp. 124-127).
5. Salameh, A. A., & Mohamed, O. (2024). Design and performance analysis of adiabatic logic circuits using FinFET technology. *Journal of VLSI Circuits and Systems*, 6(2), 84-90. <https://doi.org/10.31838/jvcs/06.02.09>
6. Sathish Kumar, T. M. (2024). Measurement and modeling of RF propagation in forested terrains for emergency communication. *National Journal of RF Circuits and Wireless Systems*, 1(2), 7-15.
7. Surendar, A. (2025). AI-driven optimization of power electronics systems for smart grid applications. *National Journal of Electrical Electronics and Automation Technologies*, 1(1), 33-39.
8. Tjoa, M. I., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Towards medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793-4813. <https://doi.org/10.1109/TNNLS.2020.3027314>
9. Yao, X., Schölkopf, B., & Zhang, Y. (2021). Interpretable deep learning for ECG classification via attention mechanism and domain knowledge. *IEEE Journal of Biomedical and Health Informatics*, 25(5), 1581-1592. <https://doi.org/10.1109/JBHI.2020.3039141>
10. Veerappan, S. (2025). Harmonic feature extraction and deep fusion networks for music genre classification. *National Journal of Speech and Audio Processing*, 1(1), 37-44.