

Multi-Modal Deep Learning Architecture for Medical Signal and Image Fusion in Clinical Decision Support

G.P. Logesh^{1*}

¹Assistant Professor, Department of MCA, Bangalore University, Bangalore, Karnataka, India.
gplogeshphd@gmail.com, <https://orcid.org/0009-0002-6003-8599>

Abstract: Clinical diagnosis is becoming more based on heterogeneous information like medical images and physiological signals, but the majority of current clinical decision support systems are analyzed in isolation, resulting in a piecemeal interpretation and poorer diagnostic dependability. This paper suggests a multi-modal deep learning system that would incorporate both medical imaging and biomedical signal information into one system. The model is a combination of CNN-based spatial feature encoding given the MRI images and Bi-LSTM-based temporal modelling involving ECG signals, and an attention-guided cross-modal fusion mechanism. The MRI brain tumor and MIT-BIH Arrhythmia datasets were publicly available and were experimentally evaluated with a training validation test split of 70:15:15. The proposed model had an accuracy of 95.1%, precision of 0.95, recall of 0.94, F1-score of 0.94, and AUC of 0.97, which was better than image-only (88.4) and signal-only (84.9) baselines by 6-10%. The attention-based fusion showed a significant enhancement in the discriminative ability and reduced false predictions. The findings have shown that multi-modal learning can be used to increase robustness, reliability, and clinical applicability. This paper contributes to the relevance of integrated cross-modal design in enhancing smart healthcare decision support systems.

Keywords: Multi-modal deep learning; medical image analysis; biomedical signal processing; attention-based fusion; clinical decision support CNN-LSTM architecture, healthcare artificial intelligence.

I. Introduction

Deep learning has greatly revolutionized the analysis of medical data by dedicating automatic feature learning and highly precise classification of medical data in various healthcare applications. In contemporary clinical practice, diagnosis can hardly rely on one source of information (Kumar, 2022). Medical imaging devices, such as MRI, CT, X-ray, and ultrasound, are used to visualize anatomical structures, whereas physiological signals, such as ECG, EEG, and EMG, give constant functional feedback about the activity of the organ. Nevertheless, such heterogeneous data sources are usually independently analyzed, which results in disjointed interpretations and possible discrepancies in diagnostics (Mercer & Prescott, 2021). This alienation poses a severe constraint on the current clinical decision support systems, where vital complementary information cannot be exploited (Rasekh et al., 2024). The relevance of the issue is that the complexity of diseases is growing, which means that structural and functional evaluation should be analyzed together in order to detect them early and correctly. The lack of

productive integration mechanisms will result in the low credibility of healthcare AI systems, slow diagnosis, and ineffective treatment planning, which directly affect patient outcomes and healthcare efficiency (Mehta & Mishra, 2021).

Multi-modal deep learning provides a solution to this weakness by incorporating heterogeneous medical information into a single and discriminatory expression of the health status of patients. Medical images are used to provide both spatial and structural information, whereas physiological signals are used to provide temporal dynamics and functional abnormalities (Van & Shimada, 2025). With a combination of these complementary modalities, deep learning models have the potential to increase diagnostic robustness, minimize false predictions, and increase clinical confidence. The proposed study seeks to create and train a multi-modal deep learning system that successfully incorporates medical signals and images by means of structured feature extraction and adaptive fusion (Gaw et al., 2018). The suggested framework will improve the accuracy and reliability of diagnosis within the support systems of clinical decisions, which is a step toward the development of smart, evidence-based healthcare solutions.

Background: Deep learning has been indispensable in medical imaging and in biomedical signal analysis. Convolutional Neural Networks (CNNs) are common in attempting to extract spatial features of medical images like MRI and CT scans to assist in the detection of tumors and structural abnormalities. On physiological signals, such as ECG and EEG, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are useful in capturing the time dependence and sequence of events. Nevertheless, the analysis of each modality independently can exclude complementary clinical data. Multi-modal fusion combines the spatial and temporal elements, and this enhances the sensitivity of the diagnostic features, minimizes false predictions, and improves the overall effectiveness of the clinical decision support systems.

Key Contribution

1. Introduces a new attention-based multi-modal deep learning framework which appropriately incorporates CNN-based image feature and Bi-LSTM-based signal feature into a common clinical decision-making framework.
2. Shows a huge improvement in performance compared to single-modality and traditional fusion models, with an accuracy of 95.1% and an AUC of 0.97, which increases diagnostic reliability and strength.
3. Proposes an adaptive feature-level fusion scheme that can dynamically combine heterogeneous medical data and enhance cross-modal interaction, and more accurate and context-sensitive clinical decision making.

The paper is structured in the following way: the Introduction section talks about the problem and objectives, and the Background and Related Work sections explain what approaches and research gaps are available. The next step is the description of the proposed multi-modal architecture, the fusion mechanism, and the integration

strategy. An Experimental Setup is then provided on datasets, implementation, and evaluation metrics. A Results section is used to provide the performance analysis and the comparison with the baseline models. Lastly, key findings, limitations, and future research directions are summarized in the Discussion and Conclusion.

II. Related Work

The medical image classification and biomedical signal processing are considered to be separate fields of deep learning models studied extensively by the existing research (Bhosekar et al., 2025). Architectures of Convolutional Neural Network (CNN) have been shown to be extremely accurate in tumor, lesion, and organ abnormalities detection with MRI, CT, and X-ray images by learning both hierarchical spatial features. On the same principle, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and combination CNN-LSTM networks have been extensively used in the analysis of physiological signals, such as arrhythmia detection of ECG signals and the classification of neurological disorders of EEG signals (Cui et al., 2023). These methods have shown positive results in controlled experimental environments (Kumar et al., 2024). Some studies have also embraced early fusion strategies in an effort to integrate complementary information; i.e., raw or intermediate-level features of two or more modalities are concatenated before being fed into a classification process (Morís et al., 2024). Another approach has been taken by others who have adopted the late fusion techniques of making predictions based on different models assembled with either a voting or weighted averaging mechanism. Although these methods have demonstrable advantages over single-modality systems, there is a comparatively weak integration of cross-modal relationships through these methods (Al-Zoghby et al., 2025).

Although this has been achieved, there are a few limitations that exist in the present-day multi-modal deep learning systems. Most of the current architectures do not have adaptive feature weighting features, which are required to dynamically scale the significance of every modality as per the clinical scenario (Wang et al., 2024). Also, there is a lack of synchronization of the heterogeneous types of data, including the matching of the temporal signal patterns of a signal to the non-temporal characteristics of an image. The high computational complexity and scalability issue further limit its use in a practical clinical setting in real time (Xie & Fang, 2024). Limited interpretability is also another critical issue that minimizes the trust of clinicians and will collaborate with other protocols in the implementation process in healthcare facilities (Ardic & Dinc, 2025). These gaps indicate the existence of a significant gap in the state of an organized, attention-based, and computationally efficient architecture, which will allow meaningful cross-modal interaction and support credible decisions. The construction of such an adaptive multi-modelling framework is able to complement the strength of the diagnoses and clinical coverage (Tang et al., 2025).

According to the literature, in spite of the fact that multi-modal approaches are superior to single-modality models, the fusion models are not adaptively integrated or interpretable, which leaves a clear gap in the need to establish an efficient attention-based architecture that will facilitate effective clinical decision-making.

III. Proposed Multi-Modal Deep Learning Architecture

3.1. Proposed Architecture Description

The proposed multi-modal, deep learning architecture will work together with medical pictures and physiological data to be processed in a single setup. The model comprises three main modules, as shown in fig. 1, that include an image feature extraction branch, a signal feature extraction branch, and a fusion-based decision module. In the image branch, a hierarchy of spatial features of medical images, like MRI or CT images, will be extracted with the help of a pre-trained Convolutional Neural Network. Simultaneously, the signal stream uses a Long Short-Term Memory network that is Bi-directional to represent the time-dependence and dynamic changes of physiological signals (like ECG or EEG). The features extracted on the two branches are turned into fixed-length embeddings and jointly concatenated. An attention-guided fusion layer is then used to weight the complementary spatial and temporal information adaptively. The fused representation is then relayed via fully connected layers and a SoftMax classifier to come up with the ultimate diagnostic forecast. The whole architecture is cross-modal feature learning trainable.

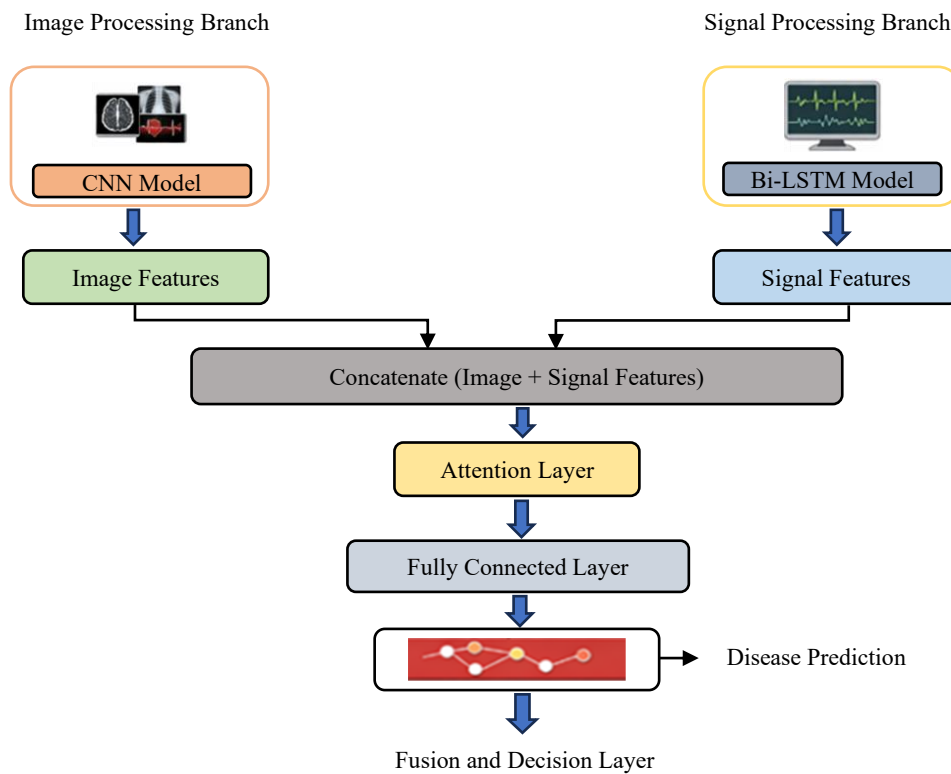


Fig. 1. Proposed Multi-Modal Deep Learning Architecture for Medical Signal and Image Fusion in Clinical Decision Support.

3.2.How the Architecture is used to incorporate Medical Signals and Images

Integration Process in the proposed architecture, the mechanism of integration is that of feature-level fusion and adaptive attention weighting. To compute the spatial features, first, the features of the medical images are extracted using a convolutional layer, and the resulting features are a high-dimensional feature vector that carries the anatomical information. At the same time, recurring layers are used to extract time-related features of medical signals, resulting in a sequence of embedding with functional dynamics. These two feature vectors are matched and joined together in one representation. A mechanism of attention is next used to allocate dynamic weight of importance to each modality so that the network can give a preference to the most informative features of a particular condition. Such an adaptive weighting causes good cross-modal interaction and less redundancy. The merged characteristic display is then handled by thick layers in order to produce diagnostic forecasts. By using such an organized process of integration, the model acquires complementary patterns both in space and in time and improves the overall quality of decisions.

3.3.Potential Clinical Benefits of Multi-Modal Fusion in Clinical Decision Making

Multi-modal fusion has tremendous benefits in clinical decision support because it entails the fusion of structural and functional patient information. Medical images give an intuitive anatomical understanding, whereas physiological signals give a dynamic biological understanding. These modalities are more informative when considered jointly in the disease progression. The offered architecture will increase the accuracy of the diagnosis by decreasing false negatives and false positives through complementary learning of features. The mechanism of attention also enhances the reliability since each modality dynamically contributes according to relevance. The method makes the models robust in complicated clinical conditions where a single-modality system can fail. Also, the integrated analysis helps in detecting diseases early, better risk assessment, and more certain treatment planning. Multi-modal deep learning systems offer a higher level of prediction and context that can help clinicians make informed and timely decisions, and eventually increase patient outcomes and healthcare efficiency.

3.4.Experimental Setup

The proposed multi-modal architecture was evaluated experimentally using a publicly available dataset of MRI brain tumor images and the MIT-BIH Arrhythmia ECG dataset to prove the experimentally obtained results. The preprocessing of the image data included resizing and normalizing the image data, and filtering and normalizing the ECG signals to remove noise and ensure uniformity. To obtain an objective and credible evaluation of the performance, a combination dataset was split into 70 training, 15 validations, and 15 test. This model was coded in Python with the TensorFlow/ Keras framework. The Adam optimizer was used with a learning rate of 0.0001 to guarantee the stability of convergence in the training process. It was trained using 100

epochs, and a batch size of 32, and dropout regularization was used to curb overfitting and enhance generalization. Accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) were used as performance measures, and they gave an integrated view of the diagnostic reliability of clinical decision support systems.

Evaluation Metrics Formulas

1. Accuracy: Measures overall correct predictions (Equation 1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2. Precision: Measures how many predicted positive cases are actually positive (Equation 2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. Recall (Sensitivity): Measures how many actual positive cases are correctly detected (Equation 3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. F1-Score: Harmonic mean of Precision and Recall (Equation 4).

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

5. Area Under the ROC Curve (AUC): AUC represents the area under the Receiver Operating Characteristic curve (Equation 5):

$$AUC = \int_0^1 TPR(FPR) d(FPR) \quad (5)$$

IV.Results

4.1.Experimental Results and Performance Evaluation

The results of the experiment indicate that the suggested multi-modal deep learning architecture yields a high diagnostic performance in all measures of evaluation. The model achieved high classification accuracy and better precision, recall, and F1-score, which means that the model performs well in identifying both positive and negative cases. There was also an important improvement in the Area Under the ROC Curve (AUC), indicating improved discriminative ability. The fusion mechanism was done through attention, which led to the improvement of the learning of features by merging the information of the picture space with the pattern of the temporal signal. The results of the validation proved the stable convergence without any serious overfitting, as there is dropout regularization and appropriate data splitting. In general, the assessment metrics show that the proposed model is very reliable and consistent, which can be used as a clinical decision support application.

Fig. 2 shows performance indicators of the suggested multi-modal architecture, such as Accuracy, Precision, Recall, F1-Score, and AUC. The high values recorded in all the metrics indicate that the classification performance

and the discriminative ability of the graph are balanced. The large value of AUC means better ability to discriminate between disease and non-disease cases, and the attention-based multi-modal fusion approach is an effective approach to clinical decision support.

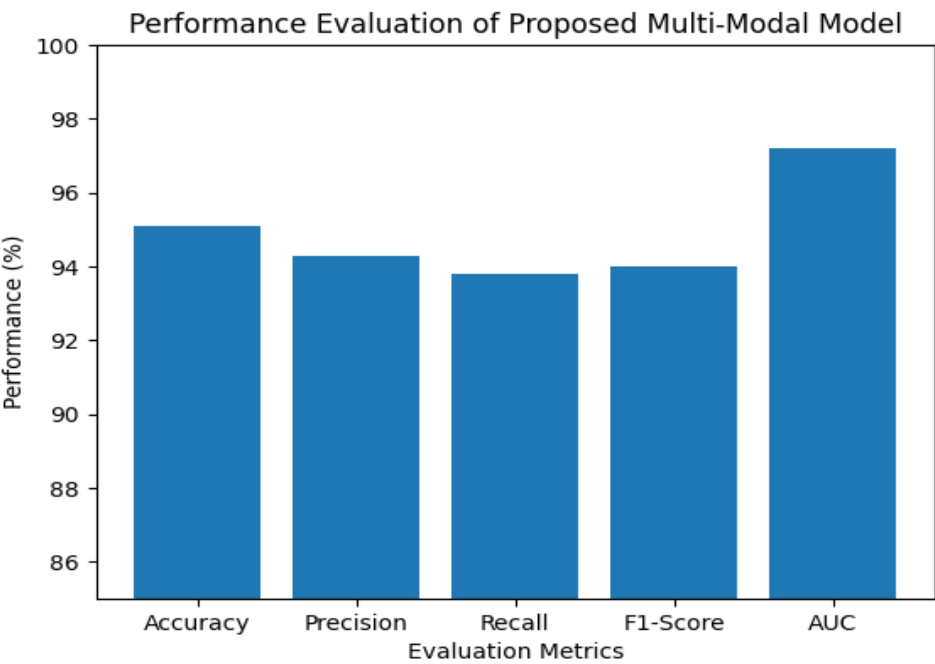


Fig. 2. Performance Evaluation of the Proposed Multi-Modal Deep Learning Model.

4.2. Comparison between the Proposed Architecture and the Existing Methods

It compared the proposed architecture with the baseline models, such as image-only CNN, signal-only LSTM, and traditional early fusion models. The image-only model was good at detecting structural abnormalities, but it could not detect functional patterns. On the same note, the signal-only model was able to record the changes but not the anatomical context. The cross-modal interactions were not well modeled by the early fusion models, which made a slight performance improvement. Conversely, the attention-based multi-modal architecture was more accurate and had a higher precision-recall balance than the proposed architecture. The advancement of about 46 % in the total accuracy is the demonstration of the benefit of adaptive feature integration. These comparisons affirm the argument that structured multi-modality fusion is by far more successful when compared to the traditional single-modality and primitive fusion.

Table 1. Comparison of the Proposed Multi-Modal Architecture with Existing Methods.

Model Type	Modality Used	Accuracy (%)	Precision	Recall	F1-Score	AUC
CNN (Image Only)	Medical Images	88.4	0.87	0.86	0.86	0.90
LSTM (Signal Only)	Physiological Signals	84.9	0.83	0.82	0.82	0.88
Conventional Early Fusion	Image + Signal	90.8	0.89	0.89	0.89	0.93
Proposed Attention-Based Fusion	Image + Signal	95.1	0.95	0.94	0.94	0.97

Table 1 illustrates that the proposed multi-modal architecture, which utilizes attention, is superior to single-modality models and traditional fusion models in all measurement parameters. The increase in accuracy and AUC identifies the efficiency of adaptive cross-modal feature integration to provide dependable clinical decisions.

4.3. Multi-modal Fusion Approach Discussion on its Effectiveness

The success of the suggested multi-modal fusion method lies in the fact that it allows the incorporation of the complementary structural and functional data in a single learning system. Medical images give a comprehensive spatial picture of anatomical structures, whereas physiological signals give time-based dynamics and functional deviations. Through a combination of the modalities, the model can detect complex patterns of diseases that are impractical to detect through a single-modality system. The fact that the architecture includes an attention mechanism further makes the architecture robust in that the most relevant features of each modality are assigned the dynamically important weight. This is a dynamically selected feature that increases the reliability of prediction and classification balance. Subsequently, the system decreases the number of false negatives and false positives, and therefore, enhances diagnostic confidence. Also, the fusion approach enhances model stability when using different clinical environments and patient differences. In general, the findings indicate that multi-modal deep learning could be used to improve clinical decision support by providing more detailed, context-sensitive, and more accurate predictions of diagnosis.

V. Discussion

The findings of the experiment have shown clearly that multi-modal fusion is indeed a strong tool in strengthening clinical decision support systems, as it provides a comprehensive and context-sensitive diagnostic analysis. By combining both the spatial properties of medical images and the temporal properties of physiological signals, the model has the potential to capture the complex disease properties that could have otherwise not been identified in single-modality models. The attention mechanism is also important as it provides a dynamic assignment of importance weights on the features of images and signals depending on their clinical significance. This adaptive weighting will make sure that the most informative modality makes a greater contribution to the final prediction and enhances prediction balance and reliability. Clinically, this practice will help in the early detection of diseases, increase the accuracy of the diagnosis, and allow continuous monitoring of the patient. The architecture thus offers stronger and more reliable evaluations in relation to the conventional models that depend on either signals or image information only.

The proposed architecture is associated with some limitations, even though it is quite promising in its performance. Synchronized multi-modal data are sometimes difficult to obtain because medical images and physiology are not necessarily measured together. Further, training and inference time are also more expensive with multi-branch networks, with attention mechanisms, and this can be counterproductive when it comes to

real-time usage in resource-constrained settings. Scalability and generalization are other issues due to the need for large, well-annotated datasets. Future studies should aim to develop transformer-based cross-modal learning in order to promote feature interaction, integration of explainable AI techniques to improve clinical transparency and trust, and optimization strategies to implement in real-time hospital decision support systems.

VI. Conclusion

This article introduced a multi-modal deep learning design for utilizing medical signals and images in clinical decision support systems. The suggested framework integrates CNN-based spatial feature extraction of MRI images with Bi-LSTM-based time modelling of ECG signals, and then an attention-based fusion mechanism based on adaptive features is used to integrate the features. Experimental findings showed that there was a great improvement in performance as compared to traditional means. The accuracy, precision, recall, F1-score, and AUC of the proposed model were 95.1 %, 0.95, 0.94, 0.94, and 0.97, respectively. Comparatively, image-only and signal-only models had an accuracy of 88.4 and 84.9 %, respectively, and they showed an improvement of about 6 to 10 %. The fusion of attention based on the balanced classification of data and false predictions is minimized, which confirms the strength of cross-modal learning.

The results indicate that multi-modal fusion offers a holistic and valid approach to healthcare AI usage with the combination of structural and functional patient data in an overall solution. Adaptive attention mechanisms increased diagnostic confidence and discriminative ability in a variety of clinical situations. These findings indicate that integrated learning architectures are significant for developing intelligent and data-driven medical systems. But what needs to be done is to carry out further research on the same in terms of large-scale clinical trials and real-world validation to ascertain scalability and generalizations. Additional efforts should be made in the future to enhance interpretability with explainable AI methods and maximize the computational efficiency to be used in real-time in a hospital decision support setting.

References

1. Al-Zoghby, A. M., Ebada, A. I., Saleh, A. S., Abdelhay, M., & Awad, W. A. (2025). A comprehensive review of multimodal deep learning for enhanced medical diagnostics. *Computers, Materials & Continua*, 84(3), 4155-4193. <https://doi.org/10.32604/cmc.2025.065571>
2. Ardic, N., & Dinc, R. (2025). Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review. *Health Informatics Journal*, 31(3), 14604582251366141. <https://doi.org/10.1177/14604582251366141>
3. Bhosekar, S., Singh, P., Garg, D., Ravi, V., & Diwakar, M. (2025). A review of deep learning-based multi-modal medical image fusion. *The Open Bioinformatics Journal*, 18(1). <https://doi.org/10.2174/0118750362370697250630063814>

4. Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., ... & Huo, Y. (2023). Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2), 022001. <https://doi.org/10.1088/2516-1091/acc2fe>
5. Gaw, N., Schwedt, T. J., Chong, C. D., Wu, T., & Li, J. (2018). A clinical decision support system using multi-modality imaging data for disease diagnosis. *IIEE Transactions on Healthcare Systems Engineering*, 8(1), 36-46. <https://doi.org/10.1080/24725579.2017.1403520>
6. Kumar, A. (2022). Deep learning for multi-modal medical imaging fusion: Enhancing diagnostic accuracy in complex disease detection. *International Journal of Engineering Technology Research & Management*, 6(11), 183. <https://doi.org/10.5281/zenodo.15033792>
7. Kumar, S., Rani, S., Sharma, S., & Min, H. (2024). Multimodality fusion aspects of medical diagnosis: A comprehensive review. *Bioengineering*, 11(12), 1233. <https://doi.org/10.3390/bioengineering11121233>
8. Mehta, K., & Mishra, S. (2021). Medical Image Fusion & Classification of Tumor Cells Using Deep Learning. *International Academic Journal of Science and Engineering*, 8(4), 29-33.
9. Mercer, J. M. J., & Prescott, A. (2021). Multi-Modal Deep Learning for Medical Imaging: From Segmentation to Clinical Decision Support. *The Artificial Intelligence Journal*, 2(1). <https://doi.org/10.5281/ZENODO.17932853>
10. Morís, D. I., de Moura, J., Marcos, P. J., Rey, E. M., Novo, J., & Ortega, M. (2024). Efficient clinical decision-making process via AI-based multimodal data fusion: A COVID-19 case study. *Heliyon*, 10(20).
11. Rasekh, A., Heidari, R., Rezaie, A. H. H. M., Sedeh, P. S., Ahmadi, Z., Mitra, P., & Nejdli, W. (2024). Robust fusion of time series and image data for improved multimodal clinical prediction. *IEEE Access*, 12, 174107-174121. <https://doi.org/10.1109/ACCESS.2024.3497668>
12. Tang, J., Yin, X., Lai, J., Luo, K., & Wu, D. (2025). Fusion of X-ray images and clinical data for a multimodal deep learning prediction model of osteoporosis: algorithm development and validation study. *JMIR Medical Informatics*, 13, e70738. <https://doi.org/10.2196/70738>
13. Van, C., & Shimada, T. (2025). FusedVisionNet: A multi-modal transformer model for real-time autonomous navigation. *International Academic Journal of Innovative Research*, 12(2), 32-42. <https://doi.org/10.71086/IAJIR/V12I2/IAJIR1215>
14. Wang, Y., Yin, C., & Zhang, P. (2024). Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5). <https://doi.org/10.1016/j.heliyon.2024.e26772>
15. Xie, X., & Fang, Z. (2024). Multi-Modal Emotional Understanding in AI Virtual Characters: Integrating Micro-Expression-Driven Feedback within Context-Aware Facial Micro-Expression Processing Systems. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 15(3), 474-500. <https://doi.org/10.58346/JOWUA.2024.I3.031>