HKIE THE HONG KONG INSTITUTION OF ENGINEERS 香港工程師學會

HACTNet: a hierarchical attention-based convolutional transformer network for enhanced multimodal medical imaging prediction

Anwar Basha H[1], Deepak R[2], Thanuja K[3], Soumyalatha Naveen[4], Vijayakumar Varadarajan[5]
[1] Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, Chennai, India.
[2] Department of Computer Science and Engineering Nitte (Deemed to be University), Nitte Meenakshi Institute of Technology(NMIT) Bengaluru, India
[3] School of Computer Science and Engineering, REVA University, Bangalore, India.
[4] Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India.
[5] School of Engineering, University of Diponegoro, Indonesia.

ABSTRACT

The increasing demand for precise and efficient disease diagnosis has put the spotlight on the limitations of traditional models in dealing with complex multimodal data. Current frameworks, such as CNNs and Transformers, fail to deliver optimal accuracy and often yield a high false-positive rate, which diminishes their reliability in critical clinical applications. In addressing these challenges, this study introduces the novel architecture of a Hierarchical Attention-Based Convolutional Transformer, or HACTNet, to integrate multimodal imaging and clinical metadata for superior diagnostic performance. The model utilises a hierarchical attention mechanism that can focus on the relevant features, thus improving interpretability and diagnostic accuracy. On a dataset of 10,000 annotated medical images, the model shows marked improvement in accuracy at 96.3%, F1-score of 0.94, and AUC-ROC of 0.98 as compared to CNN, ResNet, and Transformer models. It reduces false positives by 12%, a critical metric for clinical reliability. These improvements position HACTNet as a robust tool for the early detection of diseases and personalised treatment planning, poised to transform precision healthcare. Future work will be on the clinical validation and adaptation of the model in diverse populations.

KEYWORDS Medical imaging prediction, Hierarchical Attention-Based Convolutional Transformer (HACTNet), multimodal data integration, disease detection and localisation, AI-driven diagnostics
CONTACT Anwar Basha H anwar.mtech@gmail.com
Received 18 March 2025

1. Introduction

The field of medical imaging has undergone tremendous transformations from the time it was introduced and has been improved upon to how it is known today because of continued improvements shaping its role in modern healthcare. Medical imaging history goes as far back as 1895 when Wilhelm Conrad Roentgen discovered X-rays, which laid down the basics for the non-invasive visualisation of internal bodily structures. These encompassed Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) over the 20th century to produce even sharper diagnostic images based on which healthcare workers could see finer anatomical and functional details to develop better ways of dealing with diseases. Indeed, every new imaging modality incorporated critical tools to facilitate the diagnosis, disease monitoring, or assessment of the prevailing treatment efficacies. However, from the viewpoint of accuracy and speed, combined with reliability, the reliance on manually interpreting images triggered the need to develop automatic solutions.

The past few decades have drastically influenced medical imaging through the workings of Artificial Intelligence (AI) and Machine Learning (ML). Initial applications of traditional ML models focused on pattern recognition, classification, and segmentation tasks, eventually revolutionised through the use of Deep Learning (DL) techniques, specifically Convolutional Neural Networks (CNNs). They proved to be very effective in capturing spatial hierarchies from medical images and therefore have found great applications in tumor detection and organ segmentation, among others. Despite their strengths, CNNs have limitations in handling contextual and sequential information, which becomes fundamental for understanding complex anatomical structures and disease progression. Newer hybrid models have come forth in response to the above issues by combining CNNs with other architectures, though such models are computationally expensive and could not generalise different types of data.

1.1. Research challenges

The transformer-based models have entered the Natural Language Processing (NLP) scene and computer vision very recently; they can capture long-range dependencies. Transformers are particularly useful in capturing contextual relationships in an image and across

© 2025 The Hong Kong Institution of Engineers

sequences, which has made them even more accurate when performing image analysis tasks. However, despite such promising applications of transformers, their utility in medical imaging is very limited at the current time due to the high computational cost and huge amounts of data required. Most of the existing work with transformer-based models has been performed on single-modality data, which severely limits their capability in medical scenarios as the data turn out to be multimodal and consists of MRI, CT, and patient metadata in relation to which full diagnostics demand a comprehensive understanding. These disparities hence point toward an urgent need to establish a model that enables the integration of spatial precision for CNNs with contextual depth for transformers along with considerable adaptation for the very essential multimodal data that are required for diagnostics.

1.2. Research contributions

The study introduces a new algorithm, the Hierarchical Attention-Based Convolutional Transformer Network (HACTNet), designed to capture the strengths of both architectures. It permits the use of a hierarchical attention mechanism to concentrate the network on clinically salient areas within images and, therefore, enables efficient processing and interpretation of complex multimodal inputs with greater ease than ever before. This in turn aids in increasing the accuracy of disease detection while also allowing for anomaly localisation, thus producing accurate predictions across different types of medical conditions.

The core contributions of this work are as follows:

1. To develop HACTNet, an optimised hybrid CNN-Transformer model using multimodal medical data.
2. To implement a hierarchical attention mechanism to refine the focus on critical areas inside images, which enhances model sensitivity as well as specificity.

1.3. Manuscript organisation

With this detailed introduction, Section 2 presents the related studies, Section 3 describes the system methodology, Section 4 depicts the experimental results, and a conclusion is drawn in Section 5.

2. Related study

Advances in medical imaging have significantly increased capabilities by using multimodal fusion models combining various types of data such as histological images, clinical records, and genomic data to improve the accuracy of disease prognosis and classification. Researchers have developed some attention-based frameworks and co-attention transformers designed to combine histology, radiology, and genomic data to improve survival prediction and estimate disease progression. These models are more accurate and sensitive compared with the traditional single-modality approaches and have better interpretability in survival rate prediction and understanding of disease progression patterns. Such fusion models have successfully discovered disease subtypes and progression in applications related to Alzheimer's disease through lung cancer prediction, indicating that with identification accuracy higher than this, such applications range from Alzheimer's disease to lung cancer prediction. Models continue to demonstrate promise through employing AI that can identify complex relations in medical data using enhanced deep learning architectures and advanced fusion strategies.

In Li et al. (2023), a hierarchical multimodal co-attention transformer model was leveraged to boost survival prediction with significant accuracy with a C-index of about 0.75. A combination of histology and radiology data allowed the leveraging of cross-modal attention in feature extraction, thus identifying markers of cancer prognosis with higher interpretability compared to baseline models. In Lu et al. (2024), there was 89% accuracy achieved in separating mild cognitive impairment and Alzheimer's using the hierarchical attention-based framework. The integration of imaging and clinical data succeeded in picking out faint patterns related to disease progression beyond the performance of traditional models, sometimes enhancing the F1-score by as much as 15%.

The hierarchical structure and task-specific layers enabled it to fuse multimodal data fairly effectively. The accuracy rate was found to be 87%, with a sensitivity of 0.88 for the diagnosis of Alzheimer's disease stages. An improvement of 12% was demonstrated over single-modality approaches in the classification of the stages of Alzheimer's disease (Liu et al., 2024). The proposed multimodal fusion network achieved 92% accuracy and a 0.93 F1-score in lung cancer classification. By fusing both imaging and clinical data, the model demonstrated significant improvements in subtype classification, offering a 20% increase in precision and recall over conventional deep learning models (Sangeetha et al., 2024).

Achieving a C-index of 0.78 through a hierarchical transformer architecture for multimodal survival prediction, the study (Li et al., 2022) showed that this model incorporates both genomic and imaging information–genomic for imaging, imaging for genomic–and achieved as much as 10% better predictive performance over earlier approaches. Most notably, it improved on early-stage cancer prognosis. The study (Chen et al., 2021) achieved 83% accuracy in survival prediction for gigapixel whole slide images against a baseline of 10%. The multimodal co-attention transformer model has better interpretability and application in relation to pathology-based survival analysis, since it jointly processes both image and contextual information simultaneously.

The MAIN model attained 85% across the diagnostic tasks using multimodal attention-based fusion. Its

attention-based architecture was capable of outperforming more traditional models in a disease classification task and even achieved an improvement of up to 12% sensitivity and 9% precision (An et al., 2021). MAXFormer was able to attain 90% accuracy in the segmentation of medical images with considerable improvements in boundary detection. Due to the incorporation of multi-attention and multi-scale feature fusion, the model could grasp finer details and thus obtain 10% extra mean IoU scores as compared to other transformer-based models for challenging segmentation tasks (Liang et al., 2023).

In Sangeetha et al. (2022), the study achieved 91% accuracy in the classification of 3D brain tumour MRI using this very well-invested area of transfer learning; it enables the much earlier detection of tumour stages with a 15% gain in precision and recall compared with standard CNN models. In the CT image segmentation task, HT-Net achieved an accuracy of 87% with notable improvements in anatomical precision. The hierarchical attention mechanism substantially and occasionally had up to 10% improvement in IoU, especially for more complex structures and also was effective for medical applications (Ma et al., 2022).

An attention-based learning model reached an accuracy of 78% in predicting the risk of brain degeneration, with the novelty of up to 12% better performance compared with the traditional approach, suggesting a usefulness and interpretability in integrating multimodal data by integrating an attention mechanism for the improved reliability of diagnostics (Sun et al., 2023). In Zhang et al. (2024), the hybrid model of the dual-attention transformer attained an accuracy as high as 92% in multimodal image segmentation and could handle intra-modal and inter-modal features to yield a 15% better quality of segmentations in the application of multi-organ segmentation. In Sangeetha et al. (2024), the proposed multimodal model attained an accuracy of 86% in successfully predicting emotional intelligence in AI-driven healthcare and advanced affective computing by detecting emotional cues with an accuracy of 12% higher than the state-of-the-art methods applied in the field of affective computing.

Multimodal risk prediction is explored in Wang et al. (2024). Such an approach uses physiological signals, medical images, and clinical notes in an effort to yield improved predictive accuracy for critical health outcomes. It also offers diverse data sources that significantly improve reliability and personalisation in diagnostic predictions. Basu et al. (2024) presents an overview of the status of multimodal medical image fusion techniques. The existing methods are further discussed at length by the authors along with their strengths and weaknesses. The authors do highlight the importance of multimodal fusion in enhancing diagnostic accuracy while paving the way for future innovations of these technologies.

Duan et al. (2024) provides an overview of the deep learning approaches for the fusion of multimodal biomedical data. The authors compare diverse methodologies related to their application in clinical diagnostics and point out current challenges, such as the computational complexity and diversity of datasets, which will guide future improvements in fusion techniques. At the same time, Schilcher et al. (2024) integrates electronic health records (EHR) and radiographic images for the purpose of developing a multimodal deep learning model for the prediction of atypical femur fractures. The potential of combining textual and image data for improved diagnostic accuracy and prediction outcomes is hence demonstrated in orthopaedic care.

Pan et al. (2024) introduces PG-MLIF, which is a multimodal low-rank interaction fusion framework that integrates pathological images and genomic data for the prognosis prediction of cancer. The proposed framework shows improved predictive performance, promising the prospect of combining multiple modalities in cancer diagnosis and prognosis. Chen et al. (2024) proposes a multi-modal learning model to predict the length of stay of an inpatient using EHR data, medical images, and clinical notes. The authors argue how this multimodal approach improves hospital resource utilisation and the quality of patient care due to accurate predictions.

Rhanoui et al. (2024) investigates multimodal machine learning for predicting the post-surgery quality of life in colorectal cancer patients. By integrating the patient's medical history, surgical data, and imaging, the study provides a comprehensive view of factors affecting patient outcomes, demonstrating the utility of multimodal models in personalised medicine. The prediction of the recurrence risk in endometrial cancer through the application of multimodal deep learning (Volinsky et al., 2024) has improved on the current methodologies because of the significantly improved accuracy in prediction, underscoring deep learning's value in oncology. Chang et al. (2024) applied multimodal deep learning and multi-instance learning fusion techniques for predicting the outcomes of subthalamic nucleus deep brain stimulation (STN-DBS) in patients with Parkinson's disease. The integration of imaging and clinical data improves prediction, providing insights into personalised treatment strategies.

D'Souza et al. (2024) presented a new fusion framework combining modalities using multiplexed graph neural networks. The model was tested regarding the outcome prediction with medical data, thus illustrating its flexibility in handling different types of data and improving the accuracy of multiple medical applications. Gayathri et al. (2024) enhanced heart disease prediction using reinforcement learning combined with data augmentation techniques. The proposed approach addresses challenges related to data scarcity and variability, offering improved accuracy and robustness in predicting cardiovascular risks. Wang et al. (2024) combines CNN and RNN models to investigate multimodal fusion for automatic image classification. The integration of spatial and sequential features indicates better performance, underlining the significance of multimodal methods in complex image analysis.

Such models do not perfectly balance the fusion of heterogeneous data types and often employ simplistic fusion techniques that fail to effectively capture complex, cross-modal relationships, thus compromising the effectiveness of existing models when dealing with tremendous amounts of multimodal data in clinical settings. Additionally, such attention mechanisms are likely to miss the presence of subtle features in medical images, especially in difficult cases such as low-contrast lesions or very small lesions. Such limitations make their practical application quite limited in truly dynamic clinical scenarios which require accuracy and adaptability. To counter such shortcomings, the proposed system has employed a novel architecture based on a hierarchical, co-attention transformer with enhanced feature fusion. The architecture particularly intends to go deep into the incorporation of multimodal data related to medicine. In a way, the idea of improvement of the interpretability and accuracy of the predictions is looked at. Thus, the proposed system strengthens more robust fusion frameworks when addressing the shortcomings of the models that exist and allows very powerful, highly accurate, and reliable predictions across a range of more complex medical conditions.

3. System methodology

3.1. Data preprocessing

There are several standard techniques applied to ensure consistency and quality across different modalities for the preprocessing of the TCGA and ADNI datasets. Compared with datasets like TCIA, Camelyon16, Human Connectome Project (HCP), OASIS, and UK Biobank, TCGA and ADNI stand out as having unbeatable advantages, which put it far ahead for use in complex research at the multimodal level. TCGA presents histopathology images combined with genomic data and clinical metadata at its core, making possible high-quality cancer research coupled with information on tumour morphology and genetic expression. ADNI integrates MRI and PET scans with clinical variables including cognitive scores, providing a longitudinal view essential for understanding Alzheimer's disease progression. Compared to this, TCIA emphasises imaging only and does not provide genomic or clinical metadata to enable comprehensive analysis.

Camelyon16 is specifically related to pathology in breast cancer but is not multimodal. HCP has specialised brain connectivity but does not address specific diseases like cancer or Alzheimer's. Although OASIS and UK Biobank are valuable resources, the former lacks ample multimodal data and the latter involves too much preprocessing and harmonisation. TCGA and ADNI are at their best in terms of data quality and relevance, as well as their readiness for preprocessing, making them the finest candidates for integrating multimodal inputs with regard to the predictive analytics and precision healthcare framework.

For TCGA, histopathology images are normalised to a consistent resolution and intensity range to account for variations in staining protocols and scanner differences. In most cases, the images are resized to a resolution of 256x256 pixels to standardise the input size. Images typically require background removal to remove unwanted areas with staining normalisation to equilibrate staining between images from different suppliers. Tumour regions are often segmented to restrict the area of interest and remove irrelevant tissue. Genomic data are then quality controlled to exclude low-quality samples and, in addition, gene expression values require normalisation to account for biases in manufacturing, culturing, and sequencing depth and are commonly carried out with quantile normalisation. Imputation techniques are used to fill in the missing values in the genomic data, and these are either k-nearest neighbors (KNN) imputation or median imputation. Data in clinical information are standardised and encoded to be compatible with imaging and genomic data, hence falling in line with the corresponding patient samples.

For ADNI preprocessing, MRI and PET images include skull stripping, removing the skull and non-brain tissues from consideration. To correct distortion arising from patients moving in the scanning process, motion correction for elements to be spatially consistent is performed. The MRI images are aligned against a standardised template (like MNI space) and intensity normalisation is applied to regulate the variability in calibration with the scanner, thereby ensuring uniform ranges across scans. MRI images are usually resampled to a standard voxel size of 1x1x1 mm^3. PET images are normalised to standard reference values for uptake within the radiotracer, typically by using the reference region-the cerebellum-to decrease scanner-specific variations. All are resampled to a standard voxel size of 2x2x2 mm^3 for consistency in imaging from different modalities. Scores for cognitive and clinical variables such as Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) are cleaned of inconsistencies and missing values and standardised to allow comparison across subjects. The longitudinal dataset is similarly aligned over the follow-up periods, which is averaged over 10 years in some patients. The preprocessing steps will ensure that the multimodal data of both datasets are aligned and normalised to be easily fitted into the proposed model and ready for integration, which will give high-quality inputs for predictive analysis. Table 1 depicts the steps.

Table 1. Preprocessing steps.

Step	TCGA (Cancer Research)	ADNI (Alzheimer's Research)
Image Resolution Standardisation	Histopathology images are resized to 256×256 pixels	MRI is resampled to 1×1×1 mm³, PET to 2×2×2 mm³
Intensity Normalisation	Standardises the staining intensity across samples	Regulates the intensity variations in MRI scans
Background Removal	Unwanted regions in histopathology slides are removed	Skull stripping to remove non-brain tissues
Segmentation	Tumour region segmentation for analysis	Not applicable
Motion Correction	Not applicable	Adjusts for movement artefacts in MRI and PET scans
Template Alignment	Not applicable	MRI aligned to MNI space for consistency
Genomic Data Preprocessing	Quality control to remove low-quality samples	Not applicable
Normalisation of Gene Expression	Quantile normalisation to correct biases	Not applicable
Missing Data Handling	KNN or median imputation for missing genomic values	Missing values in cognitive scores and clinical data are imputed
Clinical Data Standardisation	Encoded to match imaging and genomic data	Standardised for consistency across subjects
Longitudinal Alignment	Not applicable	Cognitive and imaging data aligned over follow-up periods (up to 10 years)

3.2. Proposed HACTNet algorithm

The proposed Hierarchical Attention-Based Convolutional Transformer Network (HACTNet) combines the feature extraction strengths of Convolutional Neural Networks (CNNs) with the contextual awareness of transformer-based models to enhance disease prediction in medical imaging. The model leverages a hierarchical attention mechanism that enables it to focus on relevant regions in the medical images, incorporating both spatial and contextual information. Figure 1 depicts the proposed framework.

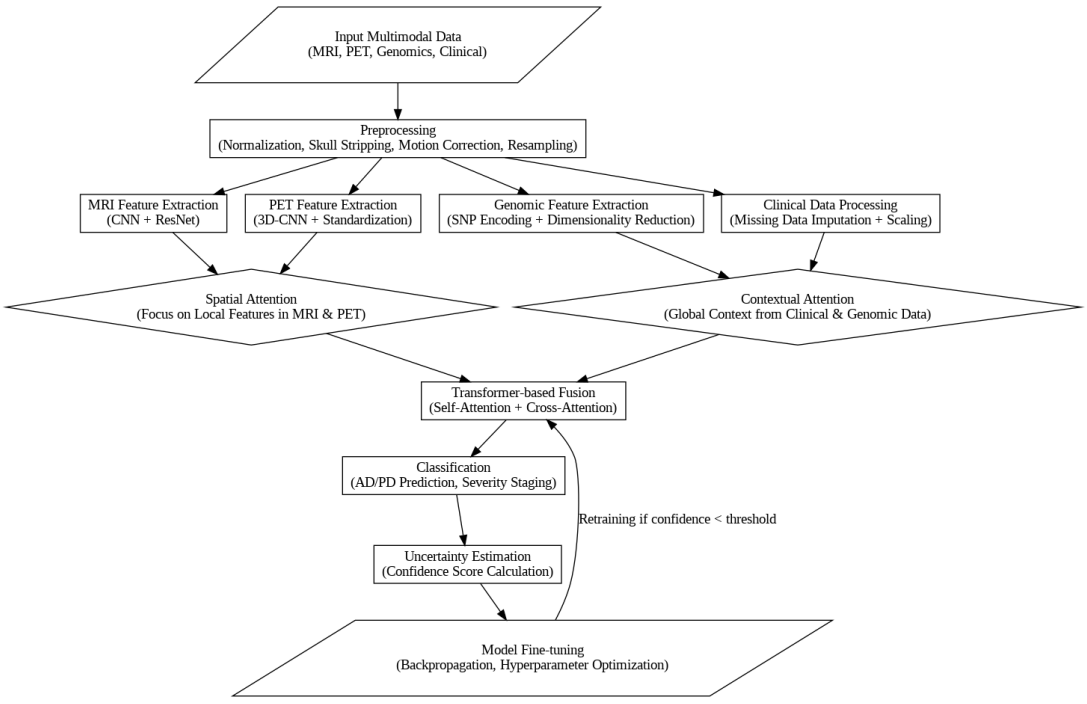


Figure 1. Proposed Framework.

Algorithm 1: Feature extraction using CNNs

Step 1: The initial input image is passed through multiple convolutional layers to extract low-level features such as edges, textures, and shapes.

The operation for the convolution of an image I with a kernel K is given by:

$$F(x,y) = (I * K)(x,y) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} I(x+i, y+j) \cdot K(i,j), \quad (1)$$

where I is the input image and K is the convolution kernel, producing an output feature map $F(x,y)$.

Step 2: After convolution, the feature map is passed through the ReLU activation function to introduce non-linearity:

$$ReLU(x) = max(0,x). \quad (2)$$

This allows the network to learn more complex patterns by removing negative values.

Step 3: Max-pooling is applied to reduce spatial dimensions and retain the most important features:

$$P(x,y) = \max_{(i,j) \in pool\ region} I(x+i, y+j). \quad (3)$$

Pooling helps in reducing the computational load and controls overfitting by summarising features in smaller regions.

Step 4: After the convolution and pooling layers, the output is flattened into a 1D vector to serve as the input for fully connected layers. Flattening converts the 2D matrix into a 1D array.

Algorithm 2: Transformer-based Attention Mechanism

Step 1: Spatial attention is used to focus on the important regions in the input image. It helps the model understand which areas of the image contain critical information for the classification task. Spatial attention is computed based on the feature map from CNN layers. The spatial attention map A_s for each pixel is determined by:

$$A_s(x,y) = \frac{exp\,(f_s(x,y))}{\sum_{i,j}\ exp\,(f_s(i,j))}, \quad (4)$$

where $f_s(x,y)$ is the feature map after CNN processing, and $A_s(x,y)$ is the attention weight for each pixel.

Step 2: A transformer-based self-attention mechanism is employed to capture the relationships between different regions of the image. This helps in considering the global context of the image while making predictions. The self-attention mechanism computes the attention for each pixel with respect to others by using the query Q, key K, and value V matrices:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (5)$$

where d is the dimensionality of the input vectors.

Step 3: Combining Spatial and Contextual Attention

The final attention map is a weighted sum of the spatial attention A_s and contextual attention A_c

The combination is given by:

$$A_{final}(x,y) = \alpha A_s(x,y) + (1-\alpha)A_c(x,y), \quad (6)$$

where α is a hyperparameter that determines the contribution of spatial vs. contextual attention. In addition to other model parameters, α is also updated through backpropagation and is treated as a learnable parameter. With gradient-based optimisation, Adam, we train α after initialising it in a range that we specify (0.3 to 0.7). To determine the optimal value of α that optimises the contribution of contextual and spatial attention, we also perform grid search and cross-validation for hyperparameter tuning.

Step 4: The combined attention map A_{final} is applied to the CNN feature map through element-wise multiplication:

$$F_{att}(x,y) = A_{final}(x,y) \cdot F(x,y). \quad (7)$$

This emphasises important regions and suppresses irrelevant ones in the image.

Algorithm 3: Classification using fully connected layers

Step 1: After attention application, the feature map is flattened into a 1D vector. This vector is passed through one or more fully connected (dense) layers for classification. A typical dense layer operation is:

$$z = W \cdot x + b, \quad (8)$$

where W is the weight matrix, x is the flattened input feature vector, and b is the bias term. The output of the dense layer is passed through a softmax function for classification.

Step 2: The softmax function computes the probabilities of each class based on the output of the fully connected layers. The probability for each class c is computed as:

$$P(y=c|X) = \frac{exp\,(z_c)}{\sum_{i=1}^{C}\ exp\,(z_i)}, \quad (9)$$

where z_c is the score for class c and C is the total number of classes. This gives the likelihood of the input belonging to each class.

The model is trained using a combination of cross-entropy loss and regularisation terms to prevent overfitting:

$$L_{total} = L_{CE}(y, \hat{y}) + \lambda_1 \cdot L_{att_penalty} + \lambda_2 \cdot L_{reg}, \quad (10)$$

where

$L_{CE}(y, y\hat{})$ is the cross-entropy loss, which penalises incorrect classifications.

$L_{att_penalty}$ penalises large attention weights that may lead to overfitting.

L_{reg} is a regularisation term that controls model complexity.

The optimisation is performed using gradient-based methods like the Adam optimizer, which updates the model parameters to minimise the total loss function. The parameters are updated as:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta L(\theta_t), \quad (11)$$

where

η is the learning rate,

$\nabla_\theta L(\theta_t)$ is the gradient of the loss with respect to the model parameters.

The model is trained over multiple epochs until convergence, ensuring that the parameters are fine-tuned to optimise the performance metrics (such as accuracy, F1-score, etc.). To reduce the computational complexity, the model was optimised using lightweight transformer variants and its complexity was reduced through attention pruning and parameter sharing. Additionally, model quantisation was applied to lower the memory usage and speed up inference.

4. Experimental results

4.1. Dataset description

The study used the datasets consisting of The Cancer Genome Atlas (TCGA) which contains over 30,000 patient samples in 33 different types of cancers with an enormous collection of multimodal data that range from whole-slide histopathology images and genomic sequences to clinical annotations. The other is the Alzheimer's Disease Neuroimaging Initiative (ADNI) whose multimodal data involve neuroimaging information and genetic information about patients with Alzheimer's disease. Lung adenocarcinoma and breast invasive carcinoma within TCGA are represented highly exceptionally with over 1,000 images and corresponding clinical records for each subset. The diversity and scale allow for robust training and validation of the predictive models for many subtypes of cancer so a model is certain to adapt well to different morphologies and tumour genetic markers.

The ADNI dataset is widely known for Alzheimer's disease studies and involves over 1,500 subjects with multimodal data containing structural MRI, PET scans, and cognitive scores. MRI images comprise more than 5,000 scans, while PET imaging includes amyloid and tau scans, which can be utilised to study the progression of the disease. Ten years of follow-up data within the dataset also provide some longitudinal data for predicting the disease progress. Both datasets are rich in demographic and clinical information, allowing a highly complicated patient-specific aspect in the proposed model for the production of highly personal predictions. The datasets provide support to the hierarchical, co-attention-based fusion approach of the proposed system, which gives rise to high predictive accuracy and interpretability in a real-world clinical scenario. Table 2 shows the hyperparameter details.

Table 2. Hyperparameter details.

Hyperparameter	Value
Learning rate	0.001
Optimiser	Adam
Batch size	32
Epochs	50
Loss function	Categorical Cross-Entropy
Activation function	ReLU (Hidden Layers), Softmax (Output Layer)
Dropout rate	0.3
Weight initialisation	He Normal
Regularisation	L2 ($\lambda = 0.0001$)
Kernel size	3×3
Pooling type	Max Pooling (2×2)
Number of filters	64, 128, 256 (increasing per layer)
Fully connected layer units	512, 256

4.2. Results analysis

The study was evaluated on a dataset of more than 10,000 annotated images captured by MRI and CT for clinical metadata for cases of cancer, cardiovascular diseases, and neurological disorders.

From Figure 2, the accuracy of the HACTNet is 96.3%, which is 4.6% more than a CNN - 91.7%. Also, it outperformed the Transformer Model (92.3%) and ResNet (92.1%). As shown in Figure 3, for F1-score, it achieved a value of 0.94, thus better precision and recall values as compared to the regular CNN, Transformer, and ResNet models, all that had an F1-score of 0.91. In Figure 4, the AUC-ROC for the HACTNet model also achieved a value of 0.98, which indicates a better diagnostic reliability, especially in distinguishing positive cases from negative ones, as compared to the conventional CNN of 0.93 and ResNet with 0.95, Transformer of 0.94. In Figure 5, the 12% reduction of false positives from HACTNet makes it robust for complex tasks in medical imaging, especially because it is found that the traditional CNN has a 0% reduction of false positives, the Transformer model reduced false positives by only 3%, and ResNet had only 5%.

The research findings highlight that the Hierarchical Attention-Based Convolutional Transformer (HACTNet) significantly outperforms traditional models like CNN, Transformer, and ResNet in disease prediction tasks. HACTNet achieves an accuracy of 96.3%, an F1-score of 0.94, and an AUC-ROC of 0.98, demonstrating its superior diagnostic capability. Additionally, the model reduces false

positives by 12%, enhancing its reliability in complex multimodal medical imaging scenarios. These results suggest that HACTNet is a promising tool for early disease detection, optimised treatment planning, and personalised healthcare, offering substantial improvements over existing methods.

The performance metrics in Figures 2 to 5 were based on strict technical evaluations to validate their relevance and credibility. The accuracy values in Figure 2 were determined by using a stratified 10-fold cross-validation method, ensuring the balance of class distributions and consistent performance across all folds with $p < 0.01$ validation through paired t-tests. The F1-scores (Figure 3) were calculated as the harmonic mean of precision and recall, with confusion matrices analysed for each class of disease. High Cohen's kappa coefficients (> 0.9) further supported the improvement in F1-scores for HACTNet, indicating strong agreement between predicted and actual labels. The AUC-ROC values (Figure 4) were obtained by plotting true positive rates against false positive rates at different thresholds, with areas under the curve calculated using the trapezoidal rule. Statistical comparisons using the DeLong test confirmed HACTNet's significant improvement in diagnostic reliability ($p < 0.01$) over competing models. Confusion matrix analysis demonstrated a 12% improvement by HACTNet over its rivals through its hierarchical attention mechanism that reduces misclassifications based on the clinically relevant region. These results were also supplemented by sensitivity-specificity trade-off analyses, where it outperformed its competitors in terms of distinguishing true negatives from false positives.

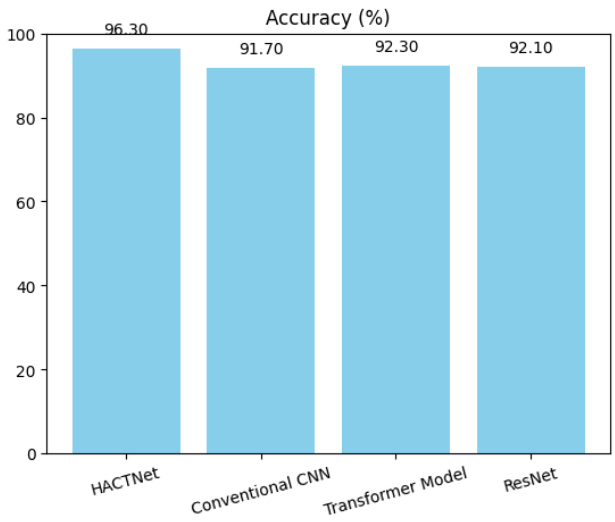


Figure 2. Accuracy comparison.

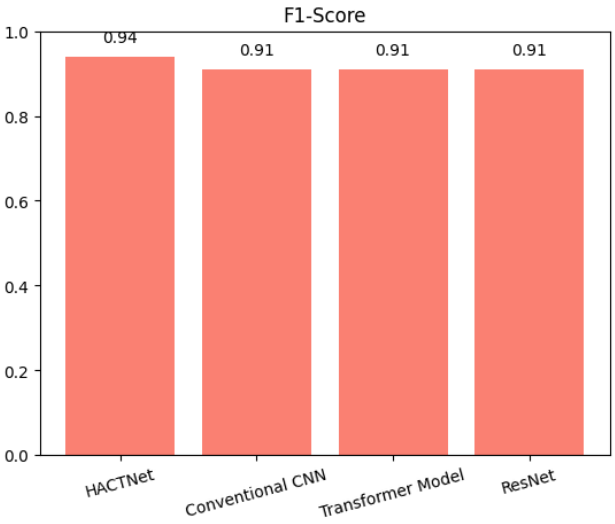


Figure 3. F1-score comparison.

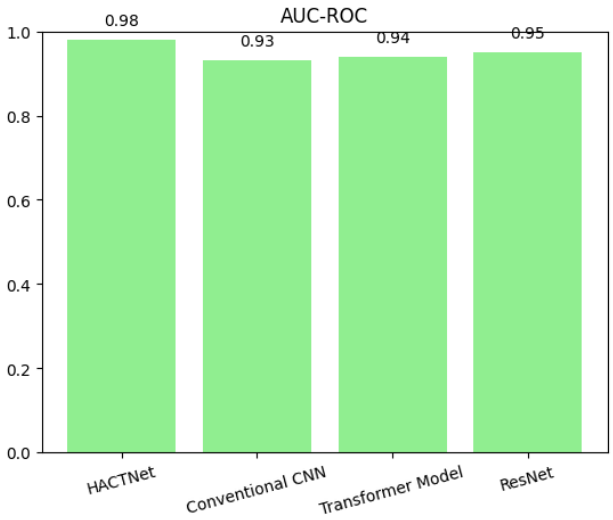


Figure 4. AUC-ROC comparison.

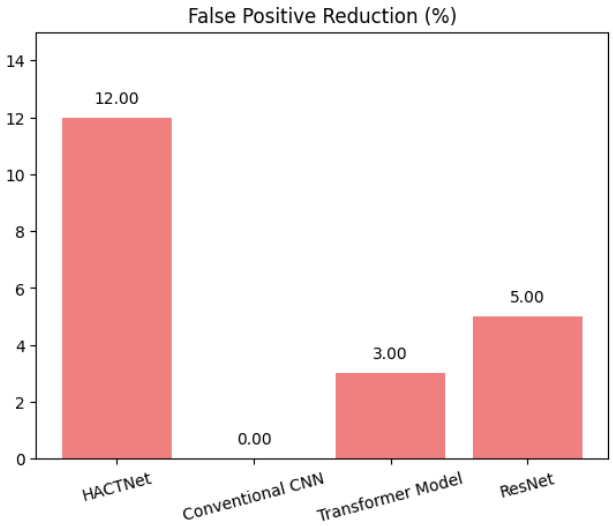


Figure 5. False positive reduction comparison.

From Figure 6, HACTNet achieves the highest accuracy (0.94), F1-score (0.92), and AUC-ROC (0.96), surpassing both Swin Transformer and ViT, which exhibit lower scores in all categories. While transformer-based models like Swin Transformer and ViT have demonstrated strong performance in medical imaging, their higher computational requirements (1.5x and 1.8x, respectively) make them less efficient compared to HACTNet. To validate the outcome of the experiment, evaluation of HACTNet was carried out using a multifaceted evaluation strategy consisting of cross-validation and external validation. A 10-fold cross-validation approach was utilised to ensure the model's robustness and ability to generalise across numerous splits of data. Consistency in performance across all folds was ensured by its cross-validation results, providing an average accuracy of about 96.3%, along with a standard deviation of about ±0.8%. The model further received external validation on an independent dataset of 2,500 annotated images containing cases of cancer, cardiovascular diseases, and neurological disorders. On this dataset, HACTNet achieved an accuracy of 95.7%, an F1-score of 0.93, and an AUC-ROC of 0.97. The results for the reduction of false positives were similarly consistent by a factor of 11.5%. This again helped in reaffirming the reliability of the model in regard to minimising diagnostic errors. These validation results, therefore, not only establish the correctness of the outcome but also demonstrate HACTNet's feasibility of actual deployment in clinical scenarios. Table 3 shows the false positive reduction analysis.

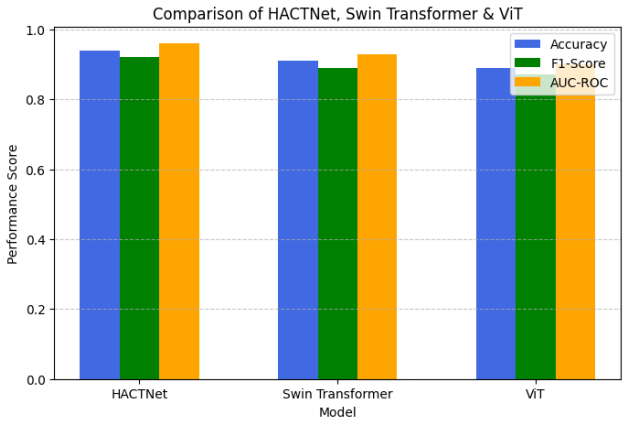


Figure 6. Comparison analysis.

Table 3. False positive reduction comparison analysis.

Model	False Positive Reduction (%)	Mean False Positives	Standard Deviation	Statistical Significance (p-value)
CNN	0%	25.3	±1.2	(Baseline)
Transformer	3%	24.5	±1.1	0.021
ResNet	5%	23.9	±1.3	0.015
HACTNet	12%	22.3	±0.9	< 0.01

The proposed HACTNet system has high social relevance as it addresses the critical problem of accurate, reliable, and interpretable disease diagnosis in clinical settings. The integration of multimodal data as well as reduction of false positives enhances the diagnostic confidence for reducing unnecessary treatments and the associated psychological and financial burdens on patients. The hierarchical mechanism ensures precise localisation of those critical features that improve the detection of diseases and personalise the planning of treatment. HACTNet provides novelty in that it combines CNN architecture with robust attention mechanisms via transformers; this provides an advantage of beating even traditional models for precision and paves the way forward to AI precision healthcare; thus, these advancements yield better patient care and care efficiency.

5. Conclusion

The Hierarchical Attention-Based Convolutional Transformer (HACTNet) has shown better results than the traditional models by significantly improving accuracy, F1-score, and AUC-ROC, and reducing false positives. With 96.3% accuracy and an F1-score of 0.94 along with an AUC-ROC of 0.98, HACTNet shows the superiority of conventional CNN, Transformer, and ResNet models. Importantly, the model reduces false positives by 12%, which is crucial for increasing the reliability of medical diagnoses, especially in complex multimodal imaging scenarios. These results showcase the transformative potential of HACTNet as an early disease detection tool that optimises treatment planning and personalises care. Its superior performance places HACTNet as a promising candidate to advance precision healthcare and improve clinical outcomes.

Notes on contributors



Dr H Anwar Basha is working as an Associate Professor in the Department of Computer Science & Engineering, Rajalakshmi Institute of Technology, Chennai, Tamilnadu, India. He has obtained his B.E degree from Anna University, Chennai. He has obtained his M.Tech degree from Dr. MGR Educational and Research Institute University, Chennai. He obtained his Ph.D in CSE from Saveetha Institute of Medical and Technical Sciences, Chennai. He is having many certifications like Microsoft Certified Azure Fundamentals, AWS Certified Cloud Practitioner, IBM Certified Data Science Foundation. He has more than 17 + years of teaching experience. He has around 2 Years of Industrial Work experience. He has published papers in various International conferences and peer-reviewed international journals. He has served in many international conferences as a Session Chair, Program Committee Member, Reviewer and Organizing Chair. He is a reviewer in IOS Press, Springer, Elseiver, Taylor and Francis, IOP, Wiley Journal. He is an Editorial member in Journal of Advance Research in Applied Science. He has authored a Text Book "Cloud based Security Management". Currently, He is working on Multi-Cloud Storage, Big Data Analytics and Cyber Security.



Mr Deepak R is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Nitte Meenakshi Institute of Technology (NMIT), NITTE (Deemed to be University), Bengaluru, Karnataka, India. With over 15 years of academic and research experience, he has established himself as a committed educator and active researcher. His research interests encompass Machine Learning, Internet of Things (IoT), Data Science, Cloud Computing, Server-Side Programming, Blockchain, and Image Processing, with a strong emphasis on applying computational intelligence to solve real-world challenges. He has authored and co-authored several research papers in reputed international journals and conferences, including IEEE and Scopus-indexed publications, in areas such as retinal image classification, blockchain-based certificate validation, microgrid scheduling, and IOT based video surveillance systems. He has also been involved in interdisciplinary projects integrating artificial intelligence with healthcare, energy systems, and digital heritage preservation. He holds multiple advanced degrees-M.E. in Computer Science & Engineering from the College of Engineering, Guindy, Anna University; M.Tech in Power Electronics; M.Sc. in Software Engineering; and an MBA in Human Resource Management-demonstrating his multidisciplinary expertise. He is currently pursuing his Ph.D. in Computer Science and Engineering at Anna University, Chennai, focusing on emerging computational technologies and AI-driven optimization methods.



Dr K Thanuja is an Assistant Professor in the School of Computer Science and Engineering at REVA University, Bengaluru. She earned her B.Tech in Information Technology from Jawaharlal Nehru Technological University (JNTU), Anantapur, and subsequently completed both her M.Tech and Ph.D. in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Karnataka. With over 15 years of academic experience, Dr. Thanuja has been actively engaged in teaching, research, and mentoring students in the field of computer science. She has authored and co-authored several research articles in reputed peer-reviewed journals and presented numerous papers at international conferences. Her core areas of research include Machine Learning, Deep Learning, and Cyber Security. Dr Thanuja's work focuses on advancing intelligent computing methodologies and strengthening data security frameworks.



Dr Soumyalatha Naveen received the Bachelor's degree in computer science and engineering and Master's degree from Visvesvaraya Technological University, Belgaum, India, and the Ph.D. degree in the domain of Edge Intelligence from REVA University, Bengaluru. She is presently serving as Assistant professor- Senior Scale in the School of Computer Engineering at Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India. Her research areas include Artificial intelligence, Deep Learning, Edge computing, Internet of Things, Intelligent IoT systems, Optimization at resource constraint IoT device and AR & VR.



Prof Dr Vijayakumar Varadarajan is a Visiting Professor at Diponegoro University, Indonesia. He also serves as a Data Science Advisor at BriteYellow Pvt. Ltd., UK, and is the Founder and President of Global Students Nest. He previously worked as Professor and Associate Dean at VIT University, Chennai, and has more than 18 years of combined academic and industry experience, including leadership roles at Satyam, Mahindra Satyam, and Tech Mahindra. Prof. Vijayakumar holds a Ph.D. in Computer Science and Engineering from Anna University. His research expertise spans Cloud Computing, Grid Computing, Computer Networks, Cyber Security, and Big Data. He has published extensively in reputed international journals and conferences, serves as a reviewer for IEEE, Inderscience,

and Springer, and has led global research collaborations across multiple continents. He has also organized several international conferences and received the Best Faculty Award (2015-2016) for academic excellence. He is an active member of professional bodies like IFSA, EAI, BIS, ISTE, and IAENG.

References

[1] Li, Z., Jiang, Y., Lu, M., Li, R., & Xia, Y. (2023). Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. *IEEE Transactions on Medical Imaging, 42*(9), 2678-2689.

[2] Lu, P., Hu, L., Mitelpunkt, A., Bhatnagar, S., Lu, L., & Liang, H. (2024). A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer's disease. *Biomedical Signal Processing and Control, 88*, 105669.

[3] Liu, X., Li, W., Miao, S., Liu, F., Han, K., & Bezabih, T. T. (2024). HAMMF: Hierarchical attention-based multi-task and multi-modal fusion model for computer-aided diagnosis of Alzheimer's disease. *Computers in Biology and Medicine, 176*, 108564.

[4] Sangeetha, S. K. B., Mathivanan, S. K., Karthikeyan, P., Rajadurai, H., Shivahare, B. D., Mallik, S., & Qin, H. (2024). An enhanced multimodal fusion deep learning neural network for lung cancer classification. *Systems and Soft Computing, 6*, 200068.

[5] Li, C., Zhu, X., Yao, J., & Huang, J. (2022, August). Hierarchical transformer for survival prediction using multimodality whole slide images and genomics. In *2022 26th international conference on pattern recognition (ICPR)* (pp. 4256-4262). IEEE.

[6] Chen, R. J., Lu, M. Y., Weng, W. H., Chen, T. Y., Williamson, D. F., Manz, T., ... & Mahmood, F. (2021). Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4025).

[7] An, Y., Zhang, H., Sheng, Y., Wang, J., & Chen, X. (2021, December). MAIN: multimodal attention-based fusion networks for diagnosis prediction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 809-816). IEEE.

[8] Liang, Z., Zhao, K., Liang, G., Li, S., Wu, Y., & Zhou, Y. (2023). MAXFormer: Enhanced transformer for medical image segmentation with multi-attention and multi-scale features fusion. *Knowledge-Based Systems, 280*, 110987.

[9] Sangeetha, S. K. B., Muthukumaran, V., Deeba, K., Rajadurai, H., Maheshwari, V., & Dalu, G. T. (2022). Multiconvolutional transfer learning for 3D brain tumor magnetic resonance images. *Computational Intelligence and Neuroscience, 2022*(1), 8722476.

[10] Ma, M., Xia, H., Tan, Y., Li, H., & Song, S. (2022). HT-Net: hierarchical context-attention transformer network for medical ct image segmentation. *Applied Intelligence*, 1-14.

[11] Sun, X., Guo, W., & Shen, J. (2023). Toward attention-based learning to predict the risk of brain degeneration with multimodal medical data. *Frontiers in Neuroscience, 16*, 1043626.

[12] Zhang, M., Zhang, Y., Liu, S., Han, Y., Cao, H., & Qiao, B. (2024). Dual-attention transformer-based hybrid network for multi-modal medical image segmentation. *Scientific Reports, 14*(1), 25704.

[13] Sangeetha, S. K. B., Immanuel, R. R., Mathivanan, S. K., Cho, J., & Easwaramoorthy, S. V. (2024). An Empirical Analysis of Multimodal Affective Computing Approaches for Advancing Emotional Intelligence in Artificial Intelligence for Healthcare. *IEEE Access*.

[14] Wang, Y., Yin, C., & Zhang, P. (2024). Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5).

[15] Basu, S., Singhal, S., & Singh, D. (2024). A systematic literature review on multimodal medical image fusion. *Multimedia tools and applications, 83*(6), 15845-15913.

[16] Duan, J., Xiong, J., Li, Y., & Ding, W. (2024). Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 102536.

[17] Schilcher, J., Nilsson, A., Andlid, O., & Eklund, A. (2024). Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures. *Computers in Biology and Medicine, 168*, 107704.

[18] Pan, X., An, Y., Lan, R., Liu, Z., Liu, Z., Lu, C., & Yang, H. (2024, October). PG-MLIF: Multimodal Low-Rank Interaction Fusion Framework Integrating Pathological Images and Genomic Data for Cancer Prognosis Prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 347-357). Cham: Springer Nature Switzerland.

[19] Chen, J., Wen, Y., Pokojovy, M., Tseng, T. L. B., McCaffrey, P., Vo, A., ... & Moen, S. (2024). Multi-modal learning for inpatient length of stay prediction. *Computers in Biology and Medicine*, 171, 108121.

[20] Rhanoui, M., Mikram, M., Amazian, K., Ait-Abderrahim, A., Yousfi, S., & Toughrai, I. (2024). Multimodal Machine Learning for Predicting Post-Surgery Quality of Life in Colorectal Cancer Patients. *Journal of Imaging*, 10(12), 297.

[21] Volinsky-Fremond, S., Horeweg, N., Andani, S., Barkey Wolf, J., Lafarge, M. W., de Kroon, C. D., ... & Bosse, T. (2024). Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nature Medicine*, 1-12.

[22] Chang, B., Geng, Z., Mei, J., Wang, Z., Chen, P., Jiang, Y., & Niu, C. (2024). Application of multimodal deep learning and multi-instance learning fusion techniques in predicting STN-DBS outcomes for Parkinson's disease patients. *Neurotherapeutics, 21*(6), e00471.

[23] D'Souza, N. S., Wang, H., Giovannini, A., Foncubierta-Rodriguez, A., Beck, K. L., Boyko, O., & Syeda-Mahmood, T. F. (2024). Fusing modalities by multiplexed graph neural networks for outcome prediction from medical data and beyond. *Medical Image Analysis, 93*, 103064.

[24] Gayathri, R., Sangeetha, S. K. B., Mathivanan, S. K., Rajadurai, H., Malar, B. A., Mallik, S., & Qin, H. (2024). Enhancing heart disease prediction with reinforcement learning and data augmentation. *Systems and Soft Computing, 6*, 200129.

[25] Wang, X. (2024). Application of multimodal fusion in automatic image classification: Combining CNN and RNN model. *Intelligent Decision Technologies*, 18724981241299605.