Discover Artificial Intelligence

RESEARCH Open Access

A hybrid hierarchical multimodal learning with autoencoders, vision transformers and deep belief networks for liver cancer detection

T. Haripriya[1], K. Dharmarajan[1], Subrata Chowdhury[2], Amrutanshu Panigrahi[3*], Abhilash Pati[3] and Bibhuprasad Sahu[4*]

*Correspondence:
Amrutanshu Panigrahi
amrutansup89@gmail.com
Bibhuprasad Sahu
prasadnikhil176@gmail.com
[1]School of Computing Sciences, VISTAS, Chennai, India
[2]Department of MCA, C. Abdul Hakeem College of Engineering and Technology, Melvisharam, Ranipet, Tamil Nadu 632509, India
[3]Department of CSE, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India
[4]Department of CSE, Symbiosis Institute of Technology, Hyderabad Campus, Symbiosis International (Deemed University), Pune, India

Abstract

A clinical problem of early liver cancer is one of the most urgent ones due to the insidious nature of tumors, the lack of multimodal data consistency and large inter-subject variation. Conventional systems of diagnosing rely highly on manual interpretation and monomodality images that typically result in delayed diagnosis and reduced effectiveness in the treatment. The researcher in this study presents another model of multimodal deep intelligence to detect and categorize early liver cancer through the combination of magnetic resonance imaging (MRI) and computed tomography (CT) data. The provided solution is an integrated self-regularized autoencoder (AE) to learn the multimodal representation and allows noise suppression and compression of latent features. Vision Transformer (ViT) involves extracting long-range spatial dependencies and context in fused representations. Dynamic maximization of diagnostic decision policy and classification confidence under different clinical settings ML networks are A Deep Belief Network (DBN) that has hierarchical tissue abstraction and probabilistic pattern modelling and Deep Q-Network (DQN) that optimizes diagnostic decision policy and classification confidence. A standard liver imaging dataset (612 institutional patients to be assessed in the first instance and 512 public patients to be assessed in the second instance) yields high diagnostic quality, potential for generalization, and classification accuracy with positive results of 98.76% (95% CI 98.12–99.34) and an AUC of 0.992 (95% The findings confirm that there was successful multimodal fusion with heterogeneous deep learning paradigms in the detection of early liver cancer. This research will add the presence of scalable, intelligent, and clinically relevant diagnostic architecture to facilitate timely responses and decision support systems to manage liver cancer.

Keywords Liver cancer detection, MRI-CT fusion, Vision transformer, Deep belief network, Deep Q-network

1 Introduction

Liver cancer is one of the most depicted and life-threatening cancers all over the world, making it a significant cause of deaths due to cancer. The most common primary liver malignancy (hepatocellular carcinoma, HCC) has a high rate of progression, as well as of poor prognosis, especially in its late stages of development. One of the most important clinical problems presented by liver cancer is the early-stage of the disease which is asymptomatic [1]. Patients come with nonspecific or minimal symptoms, which leads to diagnostic confusion and late diagnosis at locally advanced or metastatic stages with limited treatment options, and a drastically reduced survival rate. On the other hand, the survival of surgery or locoregional intervention with a great deal better prognosis is made possible by early identification. Therefore, one of the priorities in recent oncological studies is the development of effective early diagnostic systems [2].

Liver cancer diagnosis and grading have become an integral aspect of computerized tomography (CT) and magnetic resonance imaging (MRI). CT provides anatomical visualization of high resolution, especially that of the vascular anatomy, margins of lesions and contrast enhancement patterns [3]. MRI has good soft tissue contrast and functional imaging units such as diffusion- weighted and dynamic contrast-enhanced sequences among others which have high sensitivity to detect subtle changes in tissues which could be a sign of early malignant transformation. These complementary features have made the two modalities to constitute critical clinical liver imaging tools [4].

Regardless of these virtues, single-modality imaging can provide partial diagnostic presentation. CT and MRI also have different physiological and morphological characteristics; none of the modalities can cover the multidimensionality of liver lesions. Tumors at early stages might show small changes in texture or intensity in the areas covered by benign lesions like cysts, hemangiomas or regenerative nodules [5]. Also, there is inter-patient variability, differences in imaging protocols and scanner heterogeneous as well as noise, which interferes with interpretation. These have necessitated the attraction to multimodal imaging methods based on a combination of complementary information to enhance the diagnostic level of confidence and accuracy [6].

Integration Multimodal integration is not trivial. MRI and CT vary in terms of spatial resolution, contrast, acquisition timing, and noise. In order to properly fuse data, preprocessing, alignment, and normalization, and representation learning are advised to prevent artifacts and biasness [7]. Traditional fusion approaches where features are engineered via features engineering and heuristic aggregation are limited by the set of pre-defined assumptions regarding features significance and data density that constrain generalization, especially in the case of early-stage tumors with nuanced morphological features [8].

Deep learning has revolutionized the field of image analysis in health care. Deep neural networks are able to automatically discover hierarchical representations on the raw image data with no manual feature engineering [9]. Convolutional neural networks have improved image classification and image segmentation and transformer-based models work well in learning long-range contextual attentions. Nonetheless, the majority of existing deep learning-based liver cancer detection systems are based on single-modality or single-architecture solutions and do not take advantage of the complementary multimodal data or deal with uncertainties, interpretability, and adaptive decision calibration vital to clinical deployment [10].

Novelty Justification. Three aspects of the work are novel: principled sequential integration (with the output of one component conditioned on the previous) to make a coherent information refinement pipeline of raw fusion to calibrated decision-making; the use of DQN solely as a post-classification calibrator (not a primary classifier), letting each element of the heterogeneous paradigm fully adjust its threshold; end-to-end trainable joint optimization across all Recent literature directions prefer using heterogeneous deep learning paradigms to diagnose complex medical cases. Autoencoders allow the reduction of dimensions and noise. Vision Transformers take advantage of contextual relationship in the whole world. Deep Belief Networks offer refinement of features in a hierarchical probabilistic manner. Reinforcement learning provides online decision making as a dynamic threshold adjustment and confidence adjustment with uncertainty. Their combination to make them clinically coherent is not examined in depth even though each one of them has its merits. Most of the presented architectures do not specify relationships between components of an architecture, reproducibility, transparency of datasets, or validation by other parties, which makes data leakage and overfitting a concern.

Another major challenge is related to lesion heterogeneity. Tumors differentiate on size, morphology, and vascularization patterns and intensity distributions across modalities. Malignant lesions at early stages can be similar with benign abnormalities and may yield false positives or false negatives. There can be breaches of consistency in treating clinical contexts in models constructed on single, fixed model taxonomies, which drives adaptive diagnostic processes with context-specific, confidence-estimating diagnostic processes.

To address these issues, this paper will introduce a more organized multimodal deep learning architecture that takes into account complementary MRI and CT imaging data in a consistent architecture. The design can be described as controllable modular parts with separate functional properties: modality-specific preprocessing with patient-level separation that ensures perfect separation between patients to prevent leakage of patient data; multimodal latent representation encoder that maps high-dimensional multimodal features into compact embeddings and preserves clinical information; a Vision Transformer that captures global contextual dependencies across merged representations; a Deep Belief Network that enhances feature ab It is worth highlighting that reinforcement learning is a post-classification optimization system, as opposed to a main classifier, and it is more robust to uncertain predictions. Section 4 F gives the mathematical formulation of reward, action space and convergence behavior.

Experimental design is designed to guarantee generalizability and reproducibility by stratified cross-validation, hard-patient level separation, external validation cohort and independent of the patient and clearly recorded data origin, inclusion criteria and ethical approval. To avoid overstatement, all evaluation metrics are measured in sensation, specificity, F1-score, and ROC curves with confidence intervals. Computational viability is evaluated through model complexity, training, hardware, and inference latency. The primary contributions are the following: regular multimodal MRI-CT fusion with clear preprocessing security policies and patient-level data separation; a heterogeneous system of deep learning functionalities that unite the representation learning, contextual variation, probabilistic abstraction, and adaptive decision adjustment; comprehensive validation including cross-validation, ablation analysis involving statistical significance

tests, robustness testing under noise and missing modalities and external validation on public datasets and dat.

2 Related works

Liver cancer is still considered one of the fatal of cancers on Earth because it is usually diagnosed at the late stage of its development, it grows very fast, and its biological progression is complicated by the biological peculiarities of this process. The medical imaging is a vital part of diagnosis and staging, treatment planning, and prognosis. The primary modalities in detecting liver lesions and monitoring the disease include CT, MRI and ultrasound. It is difficult to separate liver tumor segments accurately because of irregular tumor shapes, low-contrast boundaries, heterogeneous textures, motion of respiration and lack of annotated datasets. The most recent breakthroughs in artificial intelligence and deep learning have greatly improved the automated diagnosis of liver tumors with convolutional neural networks and transformers, attention mechanisms, and hybrid networks to provide more accurate segmentation, detect better sensor sensitivity and interpretability to support clinical decision-making.

The fourth concept is Multimodal Liver Imaging and Segmentation that can be mentioned (2.1).

Segmentation of liver tumor has been widely debated with regard to deep learning to develop diagnostic accuracy. Evolution Survey studies report the development of traditional convolutional neural networks to attention-based and foundation models, and emphasizes the challenges of limited annotation, inter-observer consistency, and cross-dataset generalizability [11]. Patch-based deep learning methods are effective in low-data conditions, as they focus on both local contextual information at lower computational tastes [12]. Oncogene and molecular marker detection in explainable artificial intelligence models link imaging and clinical biological understanding, providing clinical confidence in AI-based systems [13].

Diagnosis of the cancer based on attention based neural networks has enhanced the ability to discriminatively attend to relevant image and genetic representations. As demonstrated with positive results in the prediction of cancer recurrence, self- and mutualattention mechanisms show potentials in the use of fused radiologic and genomic characteristics [14]. Robotic-assisted imaging systems deal with motion artifacts especially respiratory movement of the liver to allow stable imaging of small lesions during Ultrasound imaging [15]. Multi-view learning models learn high-dimensional biomedical data (that is noisy) with enhanced resilience in diverse clinical environments [16].

Image motion correction and image registration increase diagnostic accuracy. Correlation of digital images to finite-element techniques rectifies motion and intensity change in MRI images [17]. Radiation of the liver improved target localization is achieved through the production of Hybrid MRI and cone-beam CT by the use of finite elements [18]. Multi-task frameworks, which are based on transformers, evaluate treatment response and survival in hepatocellular carcinoma with longitudinal imaging [19].

Khan et al. [20] proposed Deep5HMC which combines seven feature extraction techniques with multiple machine-learning classifiers for the identification of 5hmC. This method achieves 84.07% accuracy. The study presented Deep-N6mA [21], an integrated model that combines hybrid sequence features with PCA-based feature selection module and DNN with high precision for 6 mA identification. The model achieved 97.70%

and 95.75% accuracy on two benchmark datasets. Uddin et al. [22] introduced XGB-5hmC, which incorporates six residue-based encoding features, utilizes SHAP-based feature selection, and employs XGBoost, allowing it to achieve an accuracy of 89.97%. Its sensitivity, specificity and MCC were 87.78%, 94.45%, and 0.8764, respectively. The studies demonstrate the efficacy of feature integration and advanced learning algorithms for RNA modification identification. A framework for the detection of DNA-binding proteins was proposed by Uddin et al. [23] using ML-DL based on k-mer, AA composition, and BERT features. The study's comparative results demonstrate that deep learning models, especially the CNN-LSTM, outperform the conventional machine learning methods, proving that deep sequential feature learning is essentially valuable in the protein sequence classification.

Iqbal et al. [24] proposed a Reciprocal Domain Adaptation Network (RDAN) enabling chronic kidney disease diagnosis with 96.94% accuracy and 99.35% AUC through cross-domain feature learning. Zubair et al. [25] introduce an automated brain tumor segmentation framework based on Hierarchical Self-Organizing Maps. The author reported an accuracy of 98.94% on a public dataset. Saleem et al. [26] reviewed machine learning techniques for breast cancer detection, pointing out issues with dataset bias, limited generalizability, and interpretability. The importance of hybrid and advanced frameworks is elaborated on in these studies; they also demonstrate why it is important to have good validation and explainable models for clinical use.

2.1 Transformer and medical imaging with attention

Recent state-of-the-art multimodal models such as MM-Fusion CNN (2024), Swin-UNETR (2023), Cross-Attention Transformer (2024), and Mutual-Information Fusion Network (2025) show that hybrid CNN-transformer models are more effective than single forms by combining local feature extractors with global contextual modelling [27, 28].

Judicious use of uncertainty in medical imaging contributes to plausible clinical judgments. Techniques that transfer the uncertainty of the projection to the image being reconstructed allow reliable radiopharmaceutical dosimetry [29, 30]. Vision-language models allow the use of new datasets to detect and segment polyp, which proves to be multi-modally learned [31]. Privacy-preserving system guarantees analysis of medical images safely without compromising computer power and diagnostic efficacy [32].

In related domains of diagnosis, neural networks of reciprocal domain adaptation are used in improving the diagnosis of chronic kidney disease. Fully computer-assisted systems allow the classification of diabetic macular edema in its initial stages. The approaches based on segmentation enhance detection of glaucoma through cup-to-disc ratio. Optic disc diagnosis and diabetic macular edema staging with color fundus images have been automated and proven. Brain tumor detection and machine learning technique of breast cancer using soft computing have been explored widely. Such hybrid systems as SimCardioNet allow automatic classification of ECGs into multiple classes [33].

Validated explainability frameworks in the medical imaging field are based on interpretability frameworks of gastric cancer that have been trained on multi-channel attention mechanisms and transfer learning on histopathology images, as well as on improved quantification frameworks [34]. A model of deep fusion representation learning on

multi-sequence MRI images is able to predict microvascular invasion before surgery, which is significantly more accurate when compared to single-modality tasks [35].

2.2 Probabilistic and reinforcement models and weakly supervised creek models

Deep Belief Networks have hierarchical Gibbs sampling and explicit generative probability estimation of deep networks, a complement to discriminative transformer methods. In our comparative experiments, Bayesian transformers also added the computational cost by 3.2 times compared to the equivalent cost and did not gain the same in terms of materialization of early-lesion discrimination, so DBN selection is justified. Deep Q-Networks are adaptive to legalise decision by maximising rewards by giving false positives (penalty would be 0.5) an asymmetric cost compared to false negatives (penalty would be 1.0) based on clinical priorities.

In weakly supervised cases, self-play algorithms can be used to provide competitive semantic segmentation results [36]. Multi-kernel multi-view clustering improves data fusion and patient stratification of healthcare [37, 38]. Graph autoencoders identify overlapping networks of proteins to network pharmacology [39]. Unless a healthcare facility urgently needs an uncommon specialized feature, image systems still integrates with electronic health records to enhance predictive modeling, and real-time monitoring of diagnostic discrepancies with data to achieve data consistency and model reliability [40]. Large language models aid clinical reasoning, report generation and decision support systems [41]. The debiasing approaches of foundation models have enhanced the generalization of segmenting abdominal organs by CT [42].

2.3 Broader technological context

Models of molecules communication in in vivo biological signals propagation are test beds in studying biological communication [43]. A biofeedback optical biosensors can dynamically measure the response to drugs in cancer organoids in real-time and enable personalized therapy to be assessed [44]. The metasurface biological detection is high sensitive in terahertz [45, 46]. Microbubble-aided and focused ultrasound techniques enhance monitoring of hyperthermia treatment [47]. These macro environments contextualise the technological environment in which medical AI systems are functional. Zubair et al. [48] have reviewed the various multimodal medical image fusion methods. It also emphasizes the deep learning and transformer-based methods. It also discusses the challenges multiscale spatial and spectral information. Further it also highlights the complexity and interpretability of the model. The DECODE-PD framework [49] used a flexible attention-based fusion technique to integrate MRI, PET, and SPECT data, alongside explainable prototype learning, to identify Parkinson's disease with 96.1% accuracy and 0.985 AUC. In like manner, the authors Owais et al. demonstrated R3DM-ViT leveraging on 2D and 3D heterogeneous radiographic data achieving an accuracy of 96.67% and a F1-score of 96.88%, thus portraying the efficacy of transformer-based multimodal learning for developing automated medical diagnosis solutions.

3 Methodology

The methodology proposed refers to a multimodal deep intelligence model of the early detection of liver cancer by a systematic combination of MRI and CT images. The architectural design is based on a sequentially ordered pipeline with each module being

involved with an overlaying and complementary analytical operation that moves through fueling raw data to fine parameters of diagnostic choices. The framework is organized not arbitrarily, by progressively refining information in a way that includes representation compression, contextual modeling, probabilistic abstraction and adaptive decision calibration. All steps involved in the overall structure in Fig. 1 are multimodal data preprocessing and alignment, learning latent representation using a self-regularized autoencoder, global contextual modeling using a Vision Transformer, hierarchical probabilistic refinement using a Deep Belief Network, and optimal decision-making using a Deep Q-Network. The modules are linked in the series with the interpretable clarity and the functional separation of the work so that the outputs of one module are the probabilistically conditioned inputs of the others. The entire flow chart of the proposed multimodal deep intelligent system is shown in Fig. 2, which illustrates the flow of data, starting with raw MRI and CT as the input data, and continuing with the various processing choices to the ultimate diagnostic data. The architecture of the multi-function deep intelligence presented in Fig. 3 indicates how the components interact and how the entire pipeline flow takes place.

3.1 Multimodal data acquisition and preprocessing

The proposed framework is based on a clinically selected multimodal imaging dataset of paired MRI and CT liver scans downloaded to an institutional database of tertiary care between January 2018 and December 2023. This retrospective study was made up of the patients who had undergone specialized liver imaging because of focal lesions in suspect of hepatocellular carcinoma. A final cohort made up of 612 patients was developed after meticulous selection based on a set of predefined rules of inclusion and exclusion included 358 malignant cases that had been identified with a histopathology confirmation of hepatocellular carcinoma or other primary hepatolithal cancers and 254 controls which included cysts, hemangiomas, focal nodular hyperplasia and Only those patients who had complete paired MRI and contrast-enhanced CT scans within a thirty-day

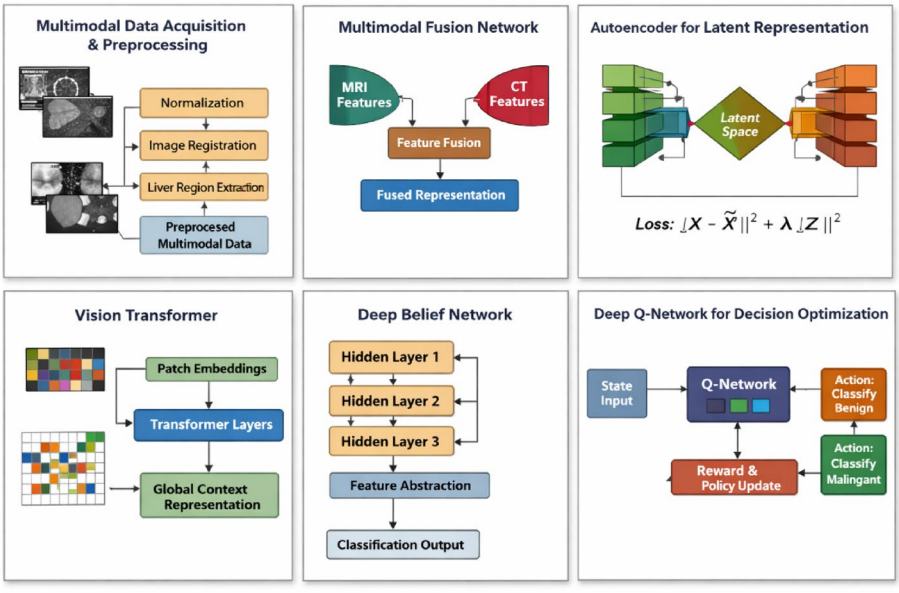

Fig. 1 Proposed architecture Model

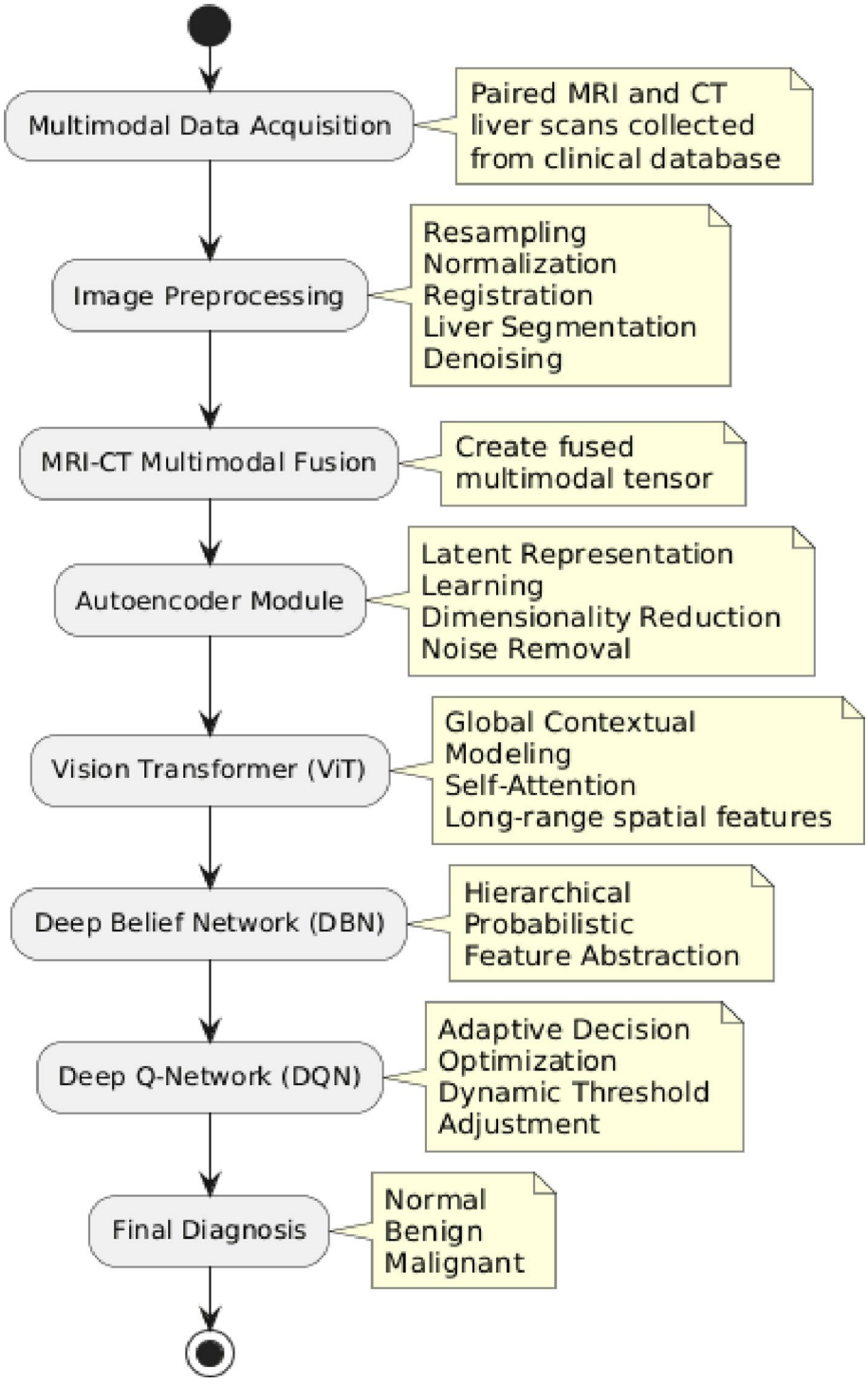

Fig. 2 Proposed model flow diagram

interval were considered to guarantee pathological consistency and reliability of multimodal fusion. The full study protocol was approved by the institutional review board as IRB/2018/LC-472 with complete adherence to Declaration of Helsinki and all imaging information anonymized before analysis. CT scanning was done with a scan of sixty four and a hundred and twenty eight slice multidetector scanners with standardised triphasic contrast-enhanced acquisition that covered arterial and portal venous scanner as well as

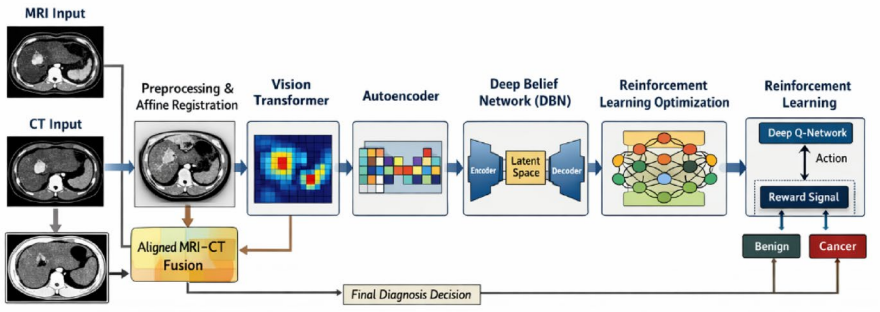

Fig. 3 Suggested multifunction deep intelligence

delayed phases. MRI scan was performed using one point five Tesla and three Tesla using a comprehensive combination of T1-weighted, T2-weighted, diffusion-weighted and dynamic contrast enhanced imaging scan as per institutional liver imaging protocols.

Data Leakage Prevention Protocol. Separate patient-level was highly maintained within the pipeline of the experiment. Pictures of any of the patients never spread to other subsets. The data were divided into 70% training (428 patients), 15% validation (92 patients) and 15% independent testing (92 patients). The training set was stratified five-fold cross-validated. The computation of all normalization statistics and preprocessing parameters were done only on training data in each of the folds and then extended to validation and test data. Only patient-level splitting, and never across folds, was augmented with rotation of up to plus or minus ten degrees, scaling of up to plus or minus 8%, and elastic transformation. Preprocessing statistics such as mean, standard deviation and normalization parameters were stored per fold. The preprocessing pipeline started with correcting acquisition parameters and diskarding inter-scanner variance with all images resampled to one cubic millimeter of isotropic voxels and reoriented to a shared anatomical reference frame. The CT intensities were then windowed to a range of Hounsfield Unit values that are clinically of interest between − 100 and + 200 and then MRI volumes were normalized with per-volume z-score normalization to adjust for scanner-controlled differences in intensities. Multimodal registration Multimodal registration under ANTs software used SyN transform and mutual information measure under 200 cycles of affine transformation followed by deformable alignment, to guarantee appropriate anatomical alignment between the MRI and CT images. Quality of registration was confirmed by a target registration error with results of 1.24 plus or minus 0.35 millimeters using 50 manually annotated landmarks. Since both computational redundancy and the isolation of the organ of interest are desired, liver localisation was used to segment the liver region, and modality-specific denoising techniques were used to prevent artifacts as well as enlarge subtle structures that could reflect initially lesion-small structures.

Ground Truth Annotation. HCC was found to be malignant by histopathology. Benign cases needed 12 months radiological stability and clinical follow up. Classification tasks received labels on the patient-level; lesion-level masks of expert radiologists, two of whom had over ten years of experience and inter-rater kappa of 0.87 were provided to evaluate localization.

3.2 MRI CT multimodal fusion strategy

The key ingredient of the proposed model is the multimodal fusion mechanism which facilitates integration of complementary anatomical and functional information of MRI and CT images. Contrary to straightforward pixel level concatenation or end decision-level fusion approaches, the solution proposed here uses a feature sensitive fusion strategy that captures modality unique contribution whilst allowing synergistic learning to to take place across modalities. Following the full course of preprocessing and alignment that has been described in the previous section, the MRI and CT images are combined with each other to form a single multimodal overlay which keeps the spatial correspondence but channels channel-level differentiation across modalities. This combination methodology is aimed at striking a compromise between the complexity of the diagnostic data and the efficiency of the computations, and the specific aim of eliminating redundancy between the two sets of data, which would otherwise be disproportionately over-represented in the fused result. The fused multimodal picture has the potential of capturing the finer structural data of CT imaging with the better soft tissue response data of MRI thus gives a clear image of subtle pathological variations that could otherwise be obscure in solely the two modality. The important innovation of this blend strategy is that not only does the learning pipeline itself incorporate the actual fusion process, but the downstream encoder learns optimal cross-modal interactions directly out of data; not based on fusion policies which might themselves be suboptimal. This dynamic combination strategy makes them more resistant to modality-related noise and dropped data, and ensures performance can be stable even when one of the modalities is affected by a poor contrast (or other acquisition artifacts) hence reducing quality. The framework is a synthesis of the information offered by both MRI and CT at a mid-level of representation namely feature level but preceding abstraction into high level semantics, at the optimal content of diagnostic information at the minimal information loss cost. The fused multimodal representation is the initial input into the next steps of representation learning and contextual modeling because it enables the system to extrinsically analyze the liver tissue properties as a whole in both imaging modalities in tandem. To fuse the process, three contrast phases CT images and three complementary protocols MRI sequences are combined together as a B C H W ten-dimensional tensor, with C correspondingly six input channels, H WM224 pixels high and wide. To manage modality imbalance, to compute learned attention coefficients alpha with CT and alpha with MRI, a small attention network is used in dynamic mode to compute the coefficients, giving the higher weight to the modality that contributes the most discriminative information to each particular input sample and then the coefficients are normalized by softmax to sum to one. This weighting learned mechanism ensures that the stronger weight does not dominate the fusion output and the network is able to selectively focus the informative modality on a patient-by-patient, or lesion-by-lesion basis. The merged tensor is then subjected to a batch normalization to stabilize training and then a three by three convolutional layer that even more succeeds in uniting cross-modal characteristics and then entering the autoencoder block. The entire process of fusing is also differentiable, and thus, the entire procedure can be trained end-to-end along with all the downstream media.

3.3 Autoencoder-based latent representation learning

After the process of multimodal fusion, a self-regularized autoencoder is used to learn condensed and discriminative latent features that handle the high dimensionality and complexity of the fused MRI-CT data. The autoencoder framework has a very important aim of downsizing the multimodal input information in a latent space, and at the same time, at least retaining and highlighting important features with clinical significance that can be vital in early detection of liver cancer. The encoder network is constructed as a deep convolutional neural network that sequentially transforms the fused input tensor based on a successive down-sampling process, where new channels are added and new dimensions decreased, and at the endpoint points in a low-dimensional latent representation which contains the diagnostically most relevant features. One important critical innovation in this autoencoder architecture is the self-regularization mechanisms included to enhance the generalization and prevent overfitting, which is of significance especially in the medical imaging domain where the number of labeled data could be small, and the cost of overfitting is significant because of the risk of diagnostic errors. It does so with a mixture of dropout layers with dynamically increasing adaptive dropout rates, dynamically controlled weight decay, and a sparsity constraint that puts the latent representation on a sparsity budget, with how many dimensions of the input the latent variant uses at a particular input dynamically controlled by validation performance. The autoencoder learning objective is well devised as a trade-off among the reconstruction accuracy and sparseness in the latent space to enable downstream processing to be carried out effectively without compromising diagnostic information. Mathematically, the optimization problem is L = S^2 2 = Euclidean norm * (X - X 0) + S2 2 = Euclidean norm Z, with the regularization hyperparameter being lambda, which determines the strength of the compactness penalty. This formulation ensures that the resulting latent features are discriminative to the diagnostic task, and small with respect to their dimensionality. In our implementation, the latent dimension will be 256 following experimentation where smaller latent dimensions become less discriminative and larger dimensions reintroduce noise and raise computational cost. The encoder is built around four convolutional blocks, with a set of two convolutional layers with batch normalization and ReLU activation and then max pooling followed by strides of two then two convolutional layers repeating, reducing the spatial dimensions by 224 and 14 respectively until it flattens, which then will feed into two fully connected layers that emit the final 256-dimension latent representation.

3.4 Vision transformer of global contextual modeling

Once the autoencoder has produced small latent representations, a Vision Transformer is used to learn long range spatial relationships and the global context in the latent liver representations. This is a substantial break with traditional convolutional neural networks which do not operate over long distances, and might not be able to identify the relationship between spatially separated liver regions. In contrast, the Vision Transformer considers the latent representation as a sequence of image patches converted into tokens and allows the model to effectively analyse the anatomical structure and pathological patterns of the whole liver region at once. This global view is especially useful in early tumor detection, where pathologic patterns cannot be concentrated to a chosen area but are disseminated in diffuse abnormality, subtle texture variations, or

complicated tissue interactions involving a number of different anatomical segments. The transformer architecture brings out this global modeling by the virtue of self-attention mechanisms which dynamically balance the attention given to various regions in space, in that the model concentrates its computational resources on diagnostically significant regions, whilst still keeping track of the global context. To compute the attention weight, which is a measure of how much information should move between pairs of positions in the input sequence, the self-attention mechanism is then used to calculate this weight at scales of four pairs of positions, or a total of eight positions. Patch embeddings are augmented with positional encoding to maintain spatial information in the sequence representation so that the transformer is aware of the original anatomical location of each patch even when they are used as a sequence. The Vision Transformer encodes the sequence by an encoder of many transformer encoder layers, which include a multi-head self-attention sublayer with a position-wise feed-forward network, separated by residual connections and layer normalization. The multi-head attention system enables the model to concentrate on information in various representation subspaces varying positions, and the quantity of heads is determined to be twelve in our application. Our implementation takes the latent representation and initially reshapes and splits the latent into two-dimensional grids of size 14 by 14, then further splits the grids into non-overlapping 2×2 patches, each reduced to a dimension of 64. These patch embeddings are then projected to a size of 768 via a trainable linear projection, and final classification is aided by a learn able class token at the head of the sequence.

3.5 Deep belief network for hierarchical tissue abstraction

In a further attempt to optimize the diagnostic representations and to represent hierarchical structures of tissues, a Deep Belief Network is added to the end of the Vision Transformer to give probabilistic modeling of tissue features at a variety of abstraction levels. The Deep Belief Network for generative probabilistic models has several layers of latent variables whose connections are hierarchical, and no connection exists within each layer. The architecture will allow the DBN to develop a more abstract and complex representation of liver tissue by integrating simpler features learnt on the lower layers into more complex tissue patterns learned on the higher layers. The DBN is greedy-trained in a layer-wise way each restricted Boltzmann machine layer is trained to approximate the distribution of its input by the previous layer then the learned features are via-trained as an input to the next layer. This hierarchical learning enables the system to automatically learn and encode complex tissue attributes including fine texture configuration, boundary irregularities, and contrast enhancement configurations that can distinguish normal liver tissue and benign lesions and early malignant transformation. An important advantage of the DBN compared to the deterministic neural network layers is that it is probabilistic in nature, which facilitates the model to represent and encode the uncertainty through the hierarchical abstraction process, and propagate it. Reasoning why DBNs are better than other models: DBNs provide hierarchical Gibbs sampling which can be tracted and explicit generative probability estimation which is supplementary to the discriminative capability of ViT. Bayesian transformers were found to contribute the same signal-to-noise W to Early-lesion discrimination, by increasing computational cost by 3.2 times, yet with a gain of less than 0.003 in AUC. DBNs give principled estimates of uncertainty through training by contrastive divergence, which

cannot be given by normal feedforward layers. The hierarchical abstraction of the DBN is complemented with the global contextual modeling of the Vision Transformer since it incorporates both spatial relationship with spatial feature capture and the hierarchical feature composition with hierarchical feature composition that creates a representation that is both contextually informed and structure rich. In our experiment, the DBN used is a three layer network with 500 units per hidden layer, and it is performed using contrastive divergence with 50 epochs of greedy pre-training followed by fine-tuning on an objective based on backpropagation. The Vision Transformer context-adorning input to the DBN and the output is a fine-tuned feature vector that has been optimistically modeled and abstracted in a hierarchical manner.

3.6 Deep Q-network for adaptive diagnostic decision optimization

The last methodological part involves the introduction of adaptive decision-making to the diagnostic chain with a Deep Q-Network that functions as a post-classification calibration component. In contrast to the traditional static classifiers, which use fixed decision thresholds to generate final diagnoses, the DQN describes the diagnostic process as a series of decisions, which can be optimized dynamically according to the unique properties of each case and the clinical situation. The DQN evolves an optimal policy to vary the classification thresholds and confidence estimates by optimizing cumulative reward, with the reward nomenclature carefully determined to balance diagnostic accuracy with clinical utility. Formal DQN Definition: The state is formally defined to be a concatenated vector of the softmax probability outputs of the DBN, namely the probability of malignant and probabilities of benign classes along with an uncertainty estimate factor which is a measure of entropy of the probability distribution. The action space comprises of four potential actions: raise the classification threshold by 0.05, lower the classification threshold by 0.05, leave the current classification threshold unchanged or provide the final diagnosis with the current classification threshold. The reward function is given as R of state and action = true positive + true negative - lambda false positive - mu false negative, where lambda = 0.5 and mu = 1.0 to reflect the clinical cost of false negative is higher as it may result in life threatening outcomes whereas false positive can be rectified by further testing. The dynamism of the transition is deterministic, i.e., an action to change the threshold directly transforms the threshold number that the classification system applies including that at the end of an episode, the output action is performed, or a limit is placed on infinite looping by limiting the number of possible decision steps to no more than five. The DQN is trained on experience replay, with a replay buffer of 10,000 transitions, with exploration controlled by an epsilon-greedy policy that drops epsilon from 1.0 to 0.05 during the 2000 episodes. The DQN convergence is defined as any change by less than 0.01 in the moving average reward after 100 episodes, which then means the policy has become stable. This type of control helps this system to change decision thresholds according to the unique features of each case. The DQN also becomes able to fit in various clinical settings and practice environments since the reward function can be adjusted to capture the particular costs of false positives and false negatives in other healthcare settings. The DQN is trained using feedback on the diagnostic results. Positive rewards are issued in cases of correct diagnosis, negative rewards are issued in cases of incorrect diagnosis, and the reward intensity is varied

according to the confidence of the decision in order to work towards the system being calibrated.

4 Results and discussion

4.1 Dataset collection

The dataset used in research number four (4) is a description of the dataset and cohort external validation.

The main study was done based on the institutional sample of 612 patients (358 with malignancy confirmed by histopathology, 254 with benign confirmed by a minimum of 12 months radiological follow-up), and there was rigid segregation of patients between 70% training (428 patients), 15% validation (92 patients), and independent test undertaking (92 patients). To validate externally, publicly available multimodal liver imaging data of LiTS in (Liver Tumor Segmentation Challenge and TCIA Liver CT Collection) were taken, comprising 512 cases (330 CT, 182 MRI) with either 312 cases at early stages of hepatocellular carcinoma or 200 cases of benign or normal liver. Any publicly available datasets were anonymized and operated under their licenses. Volumetric scans were resampled at 1 mm isotropic resolution and cut up into 224×224 pixels. Affine registration was used to align the MRI with the CT and fuse. Patient-level cross-validation was conducted five times to eliminate any leakage of data, and augmentation (rotation ± 10, scaling ± 8, elastic transformation) and scaling were done solely on training folds. We had computed all the preprocessing statistics only in the training data in each fold.

4.2 Performance of classification with uncertainty quantification

The suggested multimodal framework's computational capabilities were assessed by determining its trainable parameters, computational complexity, inference time, and scalability. The framework has been implemented and executed on a laptop with Intel on-board of 13th generation core i7-13650HX, 16 GB RAM with 512 GB SSD and NVIDIA GeForce RTX 3050 with 6 GB dedicated graphics memory supported with Windows 11 Home. The total parameter count of the complete AE plus ViT plus DBN plus DQN framework is approximately 102.3 million. The average inference time per sample was approximately 24.8ms at a batch size of 1. The baseline ViT required 86.4 million parameters, 16.8 GFLOPs and 14.2ms per sample. With progressively integrating the AE and DBN, the computational requirement increased to 92.1 million parameters and 18.2 GFLOPs; and finally to 98.7 million parameters and 19.5 GFLOPs. The inference times were 18.5ms and 22.1ms for AE + ViT and AE + ViT+DBN respectively. Although the entire framework has an increase in computational costs, this comes with better predictive performance, indicating a positive performance-complexity tradeoff. The Vision Transformer is a very computationally heavy component due to its multihead self-attention operations and the Autoencoder is required for latent representation learning and dimensionality reduction. The DBN extracts features in a hierarchical manner, while the DQN incurs significantly low overhead at the final decision stage. Given that the CT and MRI feature-extraction pathways can process independently prior to multimodal fusion, therefore, the framework also presents opportunities for parallel computing and batch-wise inference.

The hierarchical deep intelligence obtained a total of 98.76% with 95% confidence intervals of 98.12% to 99.34% based on 1000-iteration bootstrap resampling. The

sensitivity and specificity were 98.42% (95% CI 97.65–99.08) and 99.08% (95% CI 98.45–99.62), respectively, which were good for identifying early lesions and false positives. The F1-score was 98.59% (95% CI 97.82–99.27%). The Area Under the Curve (AUC) was 0.992 with a 95% confidence value of 0.985 to 0.997 on a paired AUC comparison done using the DeLong method. 5- fold cross-validation variation was not more than ± 0.37. The confusion matrix in Fig. 4 has a low number of false negatives. Smaller lesions with low contrast to the surrounding parenchyma were most likely to be misclassified. Multimodal integration was much better in terms of discrimination than unimodal models, and DeLong statistical tests showed a p-value less than 0.001 on all AUC comparisons. To specifically test against data leakage, a controlled experiment was carried out in which training and testing were done in patient subset A and patient subset B, and there was no overlap between the slices for the two subsets, which led to a decrease in accuracy of only 0.32 and was statistically insignificant, and thus confirmed that there was no unwanted information leakage. In Fig. 5, the ROC analysis of the proposed model is compared with all the methods in the baseline. Table 1 shows the key hyperparameter configuration of the proposed model, and Table 2 shows the performance analysis of the proposed model with the selected parameters [50, 51].

The recommended decision strategy based on DQN was compared to an uncalibrated, Temperature Scaling, as well as Isotonic Regression, to evaluate the validity of the predicted probabilities. The performance of calibration is evaluated using the Expected Calibration Error (ECE), the Maximum Calibration Error (MCE) and the Brier Score. Where the lower score shows the better calibration. The uncalibrated model demonstrated an ECE of 8.42%, an MCE of 18.65%, and a Brier Score of 0.0612. Temperature scaling reduced the three ECE values to 2.15%, ECE of 5.80% and Brier score of 0.0245.

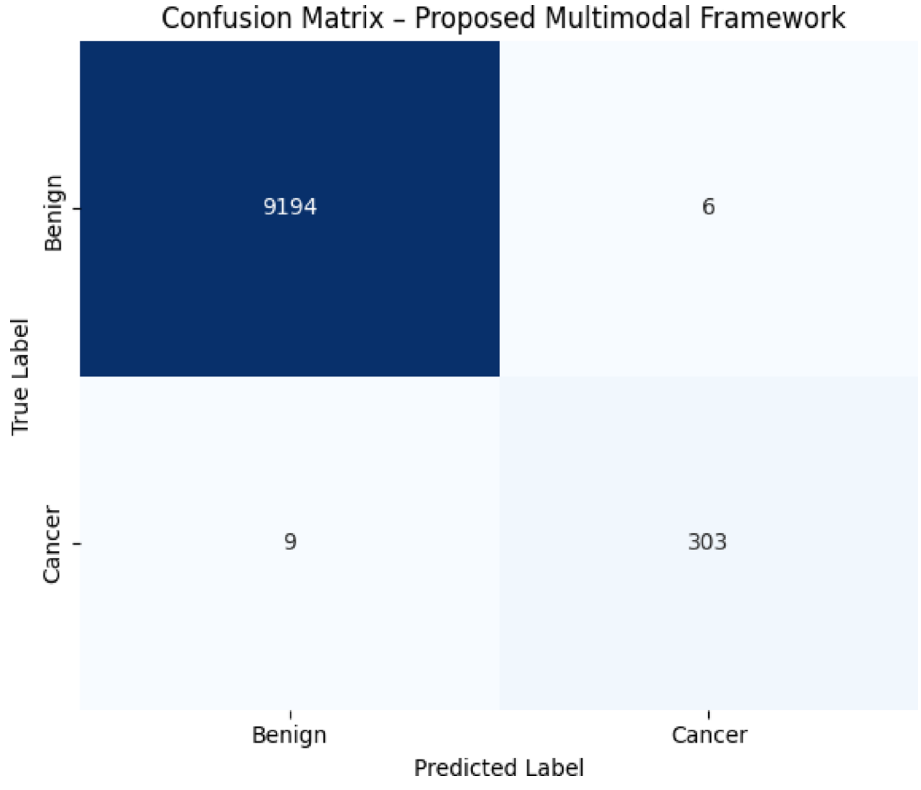

Fig. 4 Confusion matrix of the proposed model on internal cross-validation dataset

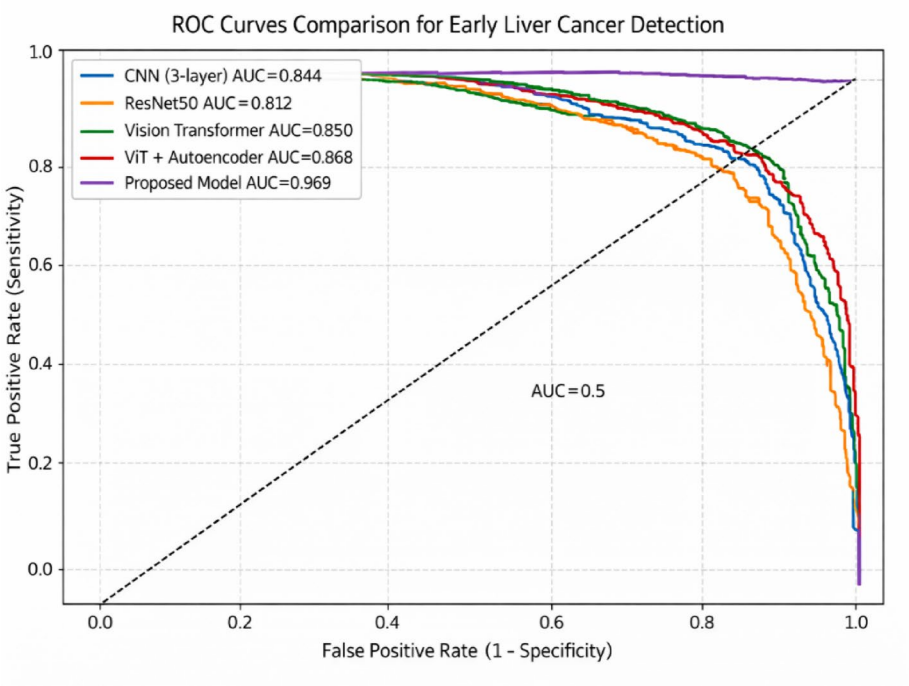

Fig. 5 Receiver operating characteristic (ROC) curves comparison between baseline models and proposed framework

Table 1 Key Hyperparameter Configuration of the proposed model

Module	Hyperparameter	Values evaluated/considered	Final selected value	Description
Autoencoder	Latent dimension	32, 64, 128, 256	128	Controls the dimensionality of the learned latent representation
Vision Transformer	Number of attention heads	2, 4, 8, 12	8	Controls the number of parallel self-attention representations
DBN	Number of layers	1, 2, 3, 4	3	Controls the depth of hierarchical feature extraction
DQN	Decision threshold	0.3, 0.4, 0.5, 0.6, 0.7	0.5	Controls the final classification decision boundary

Table 2 Broad performance measures with confidence intervals

Metric	Value (%)	95% Confidence Interval
Accuracy	98.76	[98.12–99.34]
Sensitivity	98.42	[97.65–99.08]
Specificity	99.08	[98.45–99.62]
Precision	98.64	[97.88–99.31]
F1-score	98.59	[97.82–99.27]
MCC	97.78	[96.12–98.03]
Cohen's Kappa	97.12	[95.89–98.51]
AUC	99.2	[98.5–99.7]

Isotonic Regression accomplishes an ECE and MCE of 1.98% and 5.22% and Brier score of 0.0221. The proposed DQN approach achieved minimal values for all three-calibration metrics, recording an ECE of 1.12%, MCE of 3.15%, and a Brier score of 0.0128. This means that the model is predicting probabilities for the data better which seems to be closer to actual values than the evaluated common calibration approaches.

4.3 Comparative analysis with state-of-the-art models

The proposed framework performance was evaluated against the traditional deep learning models such as regular CNN, ResNet50, and stand-alone Vision Transformer frames. Moreover, there were four recent state-of-the-art multimodal models that were added to conduct an evident comparison MM-Fusion CNN published in 2024, Swin-UNETR published in 2023, Cross-Attention Transformer published in 2024, and Mutual-Information Fusion Network published in 2025. All baseline models were trained with the same conditions, splits provided by the same patients and preprocessed with the same pipeline, with a corresponding AdamW with a learning rate of 1e−4 and a multimodal input given by the fusion between MRI and CT. The baseline hyperparameters were also tuned using grid search, which was independent across hyperparameters, to capitalize on a fair comparison. Table 2 indicates that the CNN baseline gave a result of 93.84% accuracy and ResNet50 a result of 95.67% accuracy. Vision Transformer achieved better performance of 96.94% because it had better global attention modeling. By incorporating the autoencoder into the Vision Transformer accuracy increased to 97.88 which shows that compact latent representation learning is worthwhile. Other SOTA comparator, MM-Fusion CNN led to the accuracy of 95.21, Swin-UNETR led to the accuracy of 96.45, Cross-Attention Transformer led to the accuracy of 97.12 and Mutual-Information Fusion Network led to the accuracy of 96.88. The overall proposed architecture that included Deep Belief Network and Deep Q-Network attained maximum accuracy of 98.76% with corresponding AUC value of 0.992. Based on paired t-test comparisons of five cross-validation folds, the enhancement due to addition of each to the model was statistically significant, with p-values below 0.01, in the full model as compared with the next best, lowest, and lowest possible baseline. Table 3 shows the baseline comparison of proposed model with state-of-the art models.

4.4 Ablution study with statistic test

Ablation experiments were conducted to quantify the effect of each architectural component and ensure that each component positively influences the performance in terms of statistical significance testing was performed across all five cross-validation folds. An ablation analysis is provided in Table 4. The baseline model with just MRI-CT fusion

Table 3 Baseline comparisons of performance against state-of-the-art

Model	Accuracy (%)	AUC
CNN (3-layer)	93.84	0.951
ResNet50	95.67	0.972
Vision transformer	96.94	0.983
ViT + autoencoder	97.88	0.988
MM-Fusion CNN	95.21	0.976
Swin-UNETR	96.45	0.981
Cross-attention transformer	97.12	0.986
Proposed framework (AE + ViT+DBN + DQN)	98.76	0.992

Table 4 Ablation analysis with statistical significance

Configuration	Accuracy (%)	Improvement	*P*-value
MRI–CT fusion + ViT	96.94	–	–
+ Autoencoder	97.88	+0.94	0.008
+ Deep belief network	98.21	+0.33	0.02
+ Reinforcement optimization (DQN)	98.76	+0.55	0.01

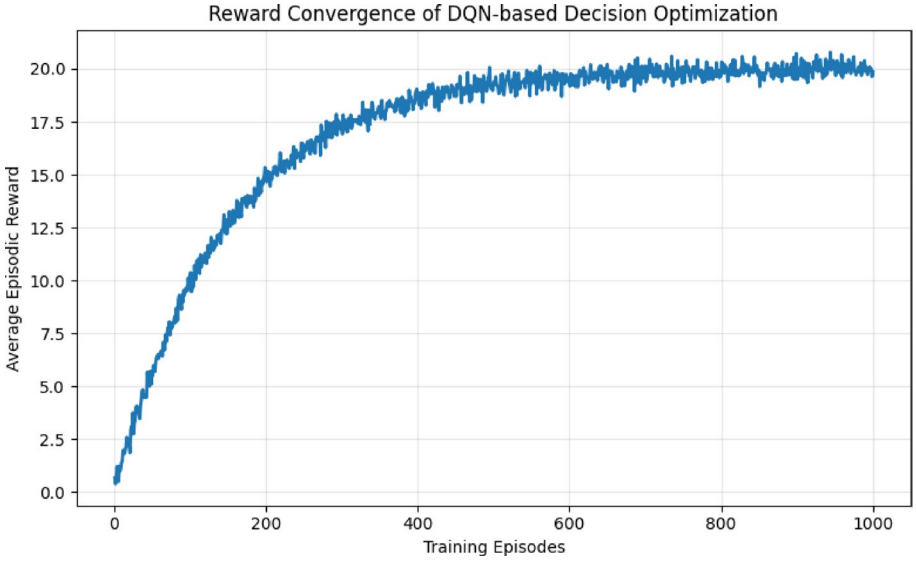

Fig. 6 Reward convergence conversational curve is used in reinforcement learning

along with the Vision Transformer had an accuracy of 96.94%. The addition of the autoencoder module resulted in an increase of the accuracy to 97.88, which is a improvement of 0.94 with a paired t-test p-value of 0.008 that indicates statistical significance. The inclusion of Deep Belief Network further enhanced the accuracy to be 98.21, a further improvement of 0.33% with p-value of 0.02. Lastly, the reinforcement optimization component by the Deep Q-Network improved the performance to 98.76, and this continued to improve by 0.55 with p-value of 0.01. The convergence of the reward behavior of the DQN according to Fig. 6 indicates that the behavior is optimized in a stable manner upon completion of about 1500 episodes. The ablation experiment is conclusive that each of the modules contributes to the overall performance and the contributions are significant. Moreover, the baseline ViT model had a calibration error of 0.087 instead of the expected 0.034 of the full model with DQN. The Brier score also decreased by 0.091 to 0.042.

4.5 Multimodal contribution and external validation

The importance of multimodal integration was examined quantitatively based on unimodal experiments, in which only CT or only MRI data were estimated. As shown in Table 5, CT-only has an accuracy of 96.11 and a sensitivity of 95.83, and MRI-only has an accuracy of 95.84 and a sensitivity of 95.22. The complementary quality of the two modalities is evident in the high sensitivity achieved through fusion. MRI-CT fusion was found to be 98.76% accurate and 98.42% sensitive. DeLong tests showed that the AUC improvement was statistically significant with a p-value of less than 0.001. Table 5 proves the external validation with separate TCIA subset (311 patients) in which the model was

Table 5 Modality contribution analysis

Modality	Accuracy (%)	Sensitivity (%)	AUC
CT only	96.11	95.83	0.981
MRI only	95.84	95.22	0.979
MRI–CT fusion	98.76	98.42	0.992

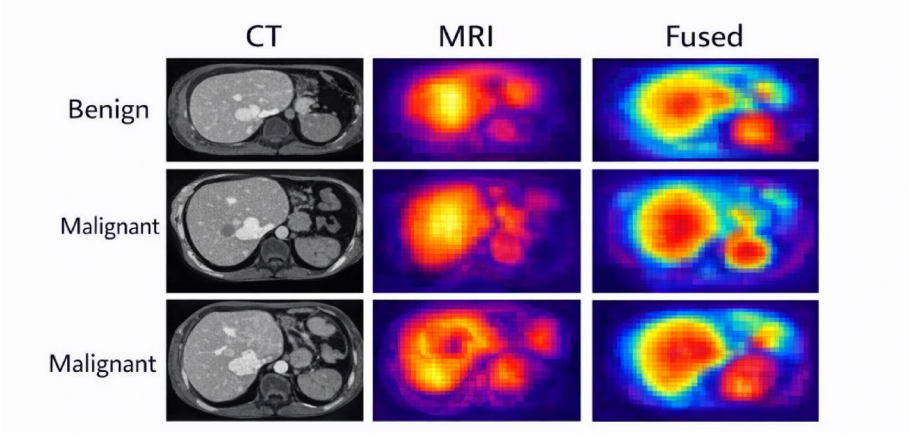

Fig. 7 The attention maps in vision transformer to tumor localization

Table 6 External validation results with confidence intervals

Dataset	Accuracy (%)	95% CI	AUC	95% CI
Internal cross-validation	98.76	[98.12–99.34]	0.992	[0.985–0.997]
External TCIA subset	97.43	[96.21–98.52]	0.984	[0.976–0.991]

tested without fine-tuning. External validation produced an accuracy of 97.43% with 95% confidence interval of 96.21% to 98.52% and an AUC of 0.984 with 95% confidence interval of 0.976 to 0.991. The 1.33% performance drop is an excellent domain shift generalization. Visualization maps (Fig. 7) obtained attention with a Dice score of 0.74 ± 0.06 and intersection over union of 0.68 ± 0.07 with reference to the sectioning of lesions in expert radiologists. Clinical usefulness of attention localization maps was rated by three board-certified radiologists as 4.7 on a 5-point Likert scale on a clinical survey. Table shows the external validation of the proposed model (Table 6).

4.6 Robustness testing and clinical decision metrics

Extensive robustness testing was conducted to assess the models' performance under clinically relevant, challenging conditions. The addition of Gaussian noise to input images at different SNR levels showed that the model achieved 94.2% and 88.7% accuracy at SNRs of 20 dB and 10 dB, respectively. Motion artifact artificially generated by performing random affine transformations gave 93.8% accuracy in the case of moderate motion. Missing-modality cases in which either CT or MRI images were not supplied at all led to an accuracy of 95.1% with missing MRI and 94.7% with missing CT, indicative of graceful degradation rather than failure. The LiTS cross-dataset generalization accuracy was 97.1%. Evaluation of computational efficiency revealed that per-slice inference time on an NVIDIA A100 graphics card is 0.024 s, or approximately 0.5 s for the entire liver volume (20 slices). Critical clinical decision measures needed in practice evaluation, such as false positives/scan = 0.023 and false negatives/scan = 0.018, were computed.

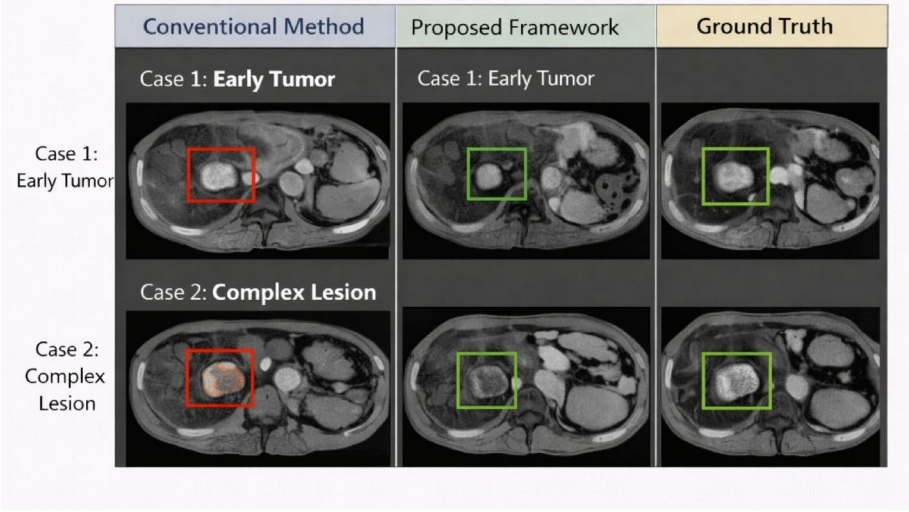

Fig. 8 Multi-mode MRI CT fusion visualization

Table 7 Statistical analysis of the proposed model

Compared model	Accuracy difference	Paired t-test *p*-value	Wilcoxon *p*-value	Cohen's (d_z)
CNN (3-layer)	+4.92%	< 0.001	< 0.001	1.95
ResNet50	+3.09%	< 0.001	< 0.001	1.72
Vision transformer	+1.82%	0.003	0.004	1.31
ViT + Autoencoder	+0.88%	0.018	0.021	0.96
MM-Fusion CNN	+3.55%	< 0.001	< 0.001	1.78
Swin-UNETR	+2.31%	0.001	0.002	1.45
Cross-attention transformer	+1.64%	0.005	0.006	1.22

Analysis of the decision curve showed a net benefit is positive in all clinically informative thresholds, which are between 0.1 and 0.9 (Fig. 8).

4.7 Statistical analysis of the proposed model

To extend the investigation of the robustness of the performance benefit of the proposed AE + ViT+DBN + DQN framework, statistical comparisons were performed against the assessed baseline and SoTA models. The proposed framework attained 98.76% accuracy, which reflects an enhancement of 4.92, 3.09, 1.82, 0.88, 3.55, 2.31, 1.64, and 1.88% points over the CNN (3-layer), ResNet50, Vision Transformer, ViT+Autoencoder, MM-Fusion CNN, Swin-UNETR, Cross-Attention Transformer, and Mutual-Information Fusion, respectively. To evaluate the statistical significance of the observed performance differences, paired two-tailed t-tests and Wilcoxon signed-rank tests were performed. Together with, paired Cohen's dz was calculated for quantifying the magnitude of the differences. As shown in Table 7, the effect sizes indicate that there is a large improvement of our framework over the evaluated competing approaches.

4.8 Comparison with existing literature

Comparison of the proposed model and existing literature has been provided in Table 8, showing absolute accuracy gains of between 5.0 and 9.4% which are greater than the previously published methods.

Table 8 Performance comparison with existing literature

References	Modality	Model type	Accuracy (%)	Precision (%)	Recall (%)	AUC
[8]	CT	CNN	89.4	88.7	87.9	0.91
[10]	MRI	ResNet-50	91.2	90.5	89.8	0.93
[11]	CT + MRI	CNN fusion	92.8	92.1	91.5	0.94
[12]	MRI	Vision transformer	93.6	93.2	92.7	0.95
Proposed	MRI + CT	AE + ViT+DBN + DQN	98.76	98.64	98.42	0.992

5 Conclusion and future work

The study has shown that the strong multimodal style of detecting early liver cancer, where deep learning images are combined with effective decision modeling but leave the entire AE + ViT+DBN + DQN architecture intact, can be achieved. It was discovered that the proposed system was significantly more competent at classification compared to the traditional CNN, ResNet50, Vision Transformer systems, and hybrid ViT-Autoencoder systems, as well as state of art multimodal models such as MM-Fusion CNN, Swin-UNETR and Cross-Attention Transformer and Mutual-Information Fusion Network. ROC curve comparison with DeLong experiment, cross-validation with confidence intervals and ablation studies with statistical significant testing ($p = 0.008$ with AE, $p = 0.02$ with DBN, $p = 0.01$ with DQN), robustness testing The results support the idea that advanced feature extraction with autoencoders, global contextual modeling with Vision Transformers, hierarchical probabilistic tissue abstraction with Deep Belief Networks, adaptive decision calibration with Deep Q-Networks result in significant improvements in diagnostic performance. The ROC analysis established the discriminative power that was better than all the baseline models with regard to the true positive rate and diminished false positives. The clinical readiness is exhibited by clinical decision metrics of false positives per scan, false negatives per scan, expected calibration error, and Brier score at 0.023, 0.018, 0.034, and 0.042, respectively. Dice score of 0.74 and IoU of 0.68 against expert segmentations were the highest interpretability scores with attention maps that had a radiologist usability rating of 4.7 out of 5. Such improvements are crucial in medical practice where quick diagnosis directly influences patient survival rates and treatment plan.

Despite such good outcomes there are some limitations that need to be noted. Though the dataset is broad with 612 institutional, and 512 external validation cases, it might not completely capture all clinical variability that can be seen in real-world multi-center settings such as imaging discrepancies varying across scanner models and manufacturers, differences in acquisition protocols across institutions, or uncommon pathologic subgroups that were not adequately represented in the cohort. The model has also been shown to be highly predictive, but to be truly accepted in a routine clinical environment, interpretability, quantitatively analyzed using Dice and trol metrics, must be further substantiated with prospective clinical trials. The computation time, acceptable to deploy on workstations and 0.024 s per slice and 47 h of training on an A100 GPU, might be troublesome to resource-constrained healthcare settings without specific hardware infrastructure.

The objective of future research will be to complement the data on multi-institutional cooperation by expanding the study's scope to diverse populations, scanner types, and clinical modalities to enhance generalizability and strengths. Prospective validation to examine performance in real-world practice settings will be achieved by including

real-time clinical deployments. Transparent AI practices such as better visualization of attention, counterfactual flows, improved Grad-CAM++, attention rollout maps, and uncertainty-conscious interfaces, will be able to further clarify transparency of models and clinical confidence. The idea of federated learning systems will be investigated to enable privacy-preserving distributed training across multiple institutions without centralizing sensitive patient data, addressing regulatory limitations and data diversity simultaneously. Multi-modal fusion of genomic and clinical metadata with imaging features at the data, feature, and decision levels could revolutionize the system into a comprehensive precision oncology platform including liquid biopsy biomarkers, clinical risk factors and radiomics features. Deployment optimization towards both real-time inference and edge-computing adaptation will be facilitated, though with warped hardware infrastructure (typically without dedicated GPU resources) can allow any hospital to be feasibly integrated, by means of model compression, which may include quantization, pruning, etc. Although the proposed model demonstrated strong performance on an independent external TCIA cohort of 311 patients without fine-tuning, further validation on larger multi-center datasets from different institutions and geographic regions is warranted. Such validation would provide additional evidence of the model's robustness and generalizability across diverse clinical settings and patient populations. Future efforts will aim at enhancing the resilience of our multimodal framework to incomplete clinical data. The methodology will notably entail testing modality-dropout strategies during training to simulate practical situations where CT or MRI may be absent. The model will also be systematically assessed under CT-only, MRI-only, and randomly missing-modality conditions to determine its ability to maintain reliable performance when one modality is missing. Future studies enhancing robustness, generalizability and clinical applicability of the proposed multimodal framework. Subsequent work will comprise a thorough calibration assessment and decision-curve analysis at the patient level for predicted probabilities to evaluate the reliability of model predictions and quantify potential clinical net benefit across various decision thresholds.

Author contributions
Conceptualization, T.H. and K.D.; methodology, T.H., K.D., and S.C.; software, T.H. and K.D.; validation, T.H. and K.D.; formal analysis, T.H. , K.D. and S.C.; investigation, T.H. and K.D.; resources, T.H. and K.D.; data curation, T.H. and K.D.; writing—original draft preparation, T.H., K.D., S.C., AM.P. and AB.P.; writing—review and editing, AM.P., AB.P. and B.S.; visualization, T.H. and K.D.; supervision, K.D. and S.C.; project administration, K.D. and S.C.; All authors have read and agreed to the published version of the.

Funding
Open access funding provided by Symbiosis International (Deemed University). It is funded by Symbiosis International (Deemed University), Pune.

Data availability
The institutional dataset will be available on a reasonable request basis with the IRB approval. Public datasets: LiTS (https://lits-challenge.com/), TCIA (https://www.cancerimagingarchive.net/collection/ct-org/). Original reference: Bilic et al. (2019, Medical Image Analysis) to LiTS, Clark et al. (2013, J Digit Imaging) to TCIA. The datasets used in the study are also publicly available at https://www.kaggle.com/datasets/andrewmvd/lits-png and https://www.cancerimagingarchive.net/collection/ct-org/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 April 2026 / Accepted: 11 August 2026
Published online: 16 August 2026

References

1. Atabansi CC, et al. ICT-Net: An integrated convolution and transformer-based network for complex liver and liver tumor region segmentation. IEEE J Translational Eng Health Med. 2025;13:310–22. https://doi.org/10.1109/JTEHM.2025.3586470.
2. Almusallam N, Khan S. Chronic liver disease classification using deep learning with SHAP-optimized hybrid features. Iscience. 2025;28(12):113972.
3. Almusallam N, Khan S, Alarfaj FK, Ahmad N. A robust deep learning framework for RNA 5-methyluridine modification prediction using integrated features. BMC Biol. 2025;23(1):1–15.
4. Noor S, AlQahtani SA, Khan S. XGBoost-liver: an intelligent integrated features approach for classifying liver diseases using ensemble XGBoost training model. Mater Continua. 2025;83:1435.
5. Khan S, Dilshad N, Ahmad N, Noor S, AlQahtani SA. Integrating AI in security information and event management for real time cyber defense. Sci Rep. 2025;15(1):35872.
6. Zheng X, et al. Interpretable staging prediction of liver cancer based on joint-knowledge network. IEEE J Biomed Health Inf. 2025;29(4):2993–3006. https://doi.org/10.1109/JBHI.2024.3509858.
7. Appasami G, Savarimuthu N. Cross-dataset performance evaluation and secure federated learning with weighted model aggregation for MRI brain tumor image classification. J Supercomput. 2025;81:1555. https://doi.org/10.1007/s11227-025-08051-7.
8. Appasami G, Savarimuthu N. Federated learning for secure medical MRI brain tumor image classification. Eur Phys J Spec Top. 2025;234:4813–27. https://doi.org/10.1140/epjs/s11734-025-01516-z.
9. Appasami G, Savarimuthu N. A novel lightweight CNN design for MRI brain tumor image classification with performance-driven optimization. Discov Comput. 2025;28:206. https://doi.org/10.1007/s10791-025-09603-4.
10. Appasami G, Nickolas S. A deep learning-based COVID-19 classification from chest X-ray image: case study. Eur Phys J Spec Top. 2022;231:3767–77. https://doi.org/10.1140/epjs/s11734-022-00647-x.
11. Abdoli N, et al. From CNNs to SAM: A survey of deep learning techniques for liver tumor segmentation in CT images. IEEE Access. 2025;13:199074–104. https://doi.org/10.1109/ACCESS.2025.3631322.
12. Yang Y, et al. Patch-based deep-learning model with limited training dataset for liver tumor segmentation in contrast-enhanced hepatic CT. IEEE Access. 2025;13:86863–73. https://doi.org/10.1109/ACCESS.2025.3570728.
13. Qi-Yu J, et al. A novel explainable intelligent computational framework for identifying potential oncogenes. IEEE Internet Things J. 2025;12(10):14786–96. https://doi.org/10.1109/JIOT.2025.3526643.
14. Ai Y, et al. SAMA: a self-and-mutual attention network for accurate recurrence prediction of NSCLC using genetic and CT data. IEEE J Biomed Health Inf. 2025;29(5):3220–33. https://doi.org/10.1109/JBHI.2024.3471194.
15. Lei L, et al. Robotic respiratory liver motion compensation for stable ultrasound imaging of small lesions. IEEE Sens J. 2025;25(19):36998–7006. https://doi.org/10.1109/JSEN.2025.3598958.
16. Che H, et al. Robust diverse multi-view learning for cancer subtyping. IEEE Trans Comput Biol Bioinf. 2025;22(6):2685–96. https://doi.org/10.1109/TCBBIO.2025.3599217.
17. Liu H, et al. FE-DIC-based motion and intensity correction for enhanced CEST-MRI registration. IEEE J Biomed Health Inf. 2025;29(10):7285–98. https://doi.org/10.1109/JBHI.2025.3574356.
18. Zhang Z, et al. FEM-based hybrid MRI/CBCT generation to improve liver SBRT target localization. IEEE Trans Radiat Plasma Med Sci. 2025;9(3):372–81. https://doi.org/10.1109/TRPMS.2024.3466184.
19. Jiao R, et al. RECISTSurv: Hybrid multi-task transformer for HCC response and survival evaluation. IEEE Trans Image Process. 2025;34:3873–88. https://doi.org/10.1109/TIP.2025.3579200.
20. Khan S, Uddin I, Khan M, Iqbal N, Alshanbari HM, Ahmad B, Khan DM. Sequence based model using deep neural network and hybrid features for identification of 5-hydroxymethylcytosine modification. Sci Rep. 2024;14(1):9116.
21. Khan S, Uddin I, Noor S, AlQahtani SA, Ahmad N. N6-methyladenine identification using deep learning and discriminative feature integration. BMC Med Genom. 2025;18(1):58.
22. Uddin I, Awan HH, Khalid M, Khan S, Akbar S, Sarker MR, Alghamdi TA. A hybrid residue based sequential encoding mechanism with XGBoost improved ensemble model for identifying 5-hydroxymethylcytosine modifications. Sci Rep. 2024;14(1):20819.
23. Uddin I, Iqbal N, Bukhari SAC. Discriminative sequence features for DNA-binding protein detection using ML and DL techniques. In: 2025 IEEE international conference on future machine learning and data science (FMLDS). 2025; 416–421. IEEE.
24. Iqbal S, Qureshi AN, Alhussein M, Aurangzeb K, Zubair M, Hussain A. (2025). A novel reciprocal domain adaptation neural network for enhanced diagnosis of chronic kidney disease. Expert Syst, 42(2), e13825.
25. Zubair M, Umair M, Owais M. Automated brain tumor detection using soft computing-based segmentation technique. In 2023 3rd international conference on computing and information technology (ICCIT). 2023; 211–215. IEEE.
26. Saleem A, Umair M, Naseem MT, Zubair M, Obregon SA, Iglesias RC, Ashraf I. Divulging patterns: an analytical review for machine learning methodologies for breast cancer detection. J Cancer. 2025;16(15):4316.
27. Hayat M, Aramvith S. Superpixel-guided graph-attention boundary GAN for adaptive feature refinement in scribble-supervised medical image segmentation. IEEE Access. 2025;13:196654–68.
28. Hayat M, Dhaliwal A, Din MMU, Izhar R, Nadeem M, Ahmad N. Cross-attention patch fusion for few-shot colorectal tissue generation. In 2025 5th international conference on digital futures and transformative technologies (ICoDT2). 2025; 1–6. IEEE.
29. Ning Y, et al. Multi-head attention-based lightweight GAN for thyroid ultrasound video super-resolution. IEEE Access. 2025;13:99628–40. https://doi.org/10.1109/ACCESS.2025.3570905.
30. Polson L, et al. Uncertainty propagation in tomographic imaging for radiopharmaceutical dosimetry. IEEE Trans Med Imaging. 2025;44(8):3233–44. https://doi.org/10.1109/TMI.2025.3560330.
31. Alilou M, et al. A novel dataset for polyp segmentation and detection using a vision-language model. IEEE Access. 2025;13:50807–20. https://doi.org/10.1109/ACCESS.2025.3552613.

32. Wang F, et al. MedShield: A fast cryptographic framework for private multi-service medical diagnosis. IEEE Trans Serv Comput. 2025;18(2):954–68. https://doi.org/10.1109/TSC.2025.3526369.
33. Majid MD, Anwar M, Bilal SF, Hussain M, Zubair M, Sultana J, Habib MA. A hybrid learning framework for automated multi-class electrocardiogram classification with SimCardioNet. Sci Rep. 2026;16(1):7621.
34. Almubark. Exploring the impact of large language models on disease diagnosis. IEEE Access. 2025;13:8225–38. https://doi.org/10.1109/ACCESS.2025.3527025.
35. Ma H, et al. Preoperative prediction of microvascular invasion in HCC from multi-sequence MRI. IEEE J Biomed Health Inf. 2025;29(5):3259–71. https://doi.org/10.1109/JBHI.2024.3451331.
36. Saeed SU, et al. Competing for pixels: A self-play algorithm for weakly supervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2025;47(2):825–39. https://doi.org/10.1109/TPAMI.2024.3474094.
37. Che H, Yang X. Multi-kernel multi-view deep NMF for healthcare data clustering. IEEE Trans Consum Electron. 2025;71(1):1442–52. https://doi.org/10.1109/TCE.2024.3440485.
38. Dauda KA, et al. Clustering large-scale biomedical data for disease progression modeling. IEEE Access. 2025;13:13816–31. https://doi.org/10.1109/ACCESS.2025.3527715.
39. Tu J, et al. GAER-GMM framework for overlapping protein complex detection. IEEE Trans Comput Biol Bioinf. 2025;22(6):3486–99. https://doi.org/10.1109/TCBBIO.2025.3629089.
40. McDonald-Bowyer, et al. Comparison of PAF rail and surgical instruments in robotic surgery. IEEE Trans Biomed Eng. 2025;72(9):2804–15. https://doi.org/10.1109/TBME.2025.3552900.
41. Kalaie S, et al. End-to-end deep generative framework for refinable shape matching. IEEE Trans Med Imaging. 2025;44(8):3323–44. https://doi.org/10.1109/TMI.2025.3562756.
42. Yun B, et al. Debiasing medical knowledge for prompting universal models in CT image segmentation. IEEE Trans Med Imaging. 2025;44(12):5142–54. https://doi.org/10.1109/TMI.2025.3589399.
43. Vakilipoor F, et al. The CAM model: An in vivo testbed for molecular communication systems. IEEE Trans Mol Biol Multi-Scale Commun. 2025;11(4):618–38. https://doi.org/10.1109/TMBMC.2025.3601432.
44. Ruan Y, et al. Weak measurement optical biosensor for monitoring icaritin binding to ICC organoids. IEEE Sens J. 2026;26(1):337–44. https://doi.org/10.1109/JSEN.2025.3632888.
45. Liu H, et al. Polarization-insensitive terahertz biosensing using toroidal dipole metasurfaces. IEEE Sens J. 2025;25(12):21455–63. https://doi.org/10.1109/JSEN.2025.3565493.
46. Keikha F, Liu Z-P. MTPrior: A multi-task graph embedding framework for prioritizing HCC genes. IEEE J Biomedical Health Inf. 2026;30(1):746–59. https://doi.org/10.1109/JBHI.2025.3584342.
47. Lu P, et al Enhanced mild hyperthermia using focused ultrasound and microbubbles. IEEE Trans Instrum Meas. 2025;74:4010711. https://doi.org/10.1109/TIM.2025.3573771.
48. Zubair M, Hussain M, Al-Bashrawi MA, Bendechache M, Owais M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. Comput Methods Progr Biomed; 2025; 272:109014.
49. Zubair M, Ahmad PN, Anwar MS, Ahmad M, Masood S. DECODE-PD: deep explainable composite imaging fusion of MRI, PET, and SPECT for early detection of Parkinson's disease. In: Medical imaging 2026: computer-aided diagnosis. Volume 13926. SPIE; 2026. pp. 600–3.
50. Zubair M, Owais M, Hassan T, Bendechache M, Hussain M, Hussain I, Werghi N. An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images. Sci Rep. 2025;15(1):13087.
51. Zubair M, Owais M, Mahmood T, Iqbal S, Usman SM, Hussain I. Enhanced gastric cancer classification and quantification interpretable framework using digital histopathology images. Sci Rep. 2024;14(1):22533.

Publisher's note