Deep Pattern Recognition Framework for Early Tumor Detection in Multimodal Medical Imaging

Nasser Al Riyami and Huda Al Kaabi

Department of Artificial Intelligence
School of Computing and Smart Technologies, Emirates International University, UAE
Department of Computer Science
College of Information Technology, Gulf Innovation University, UAE

ABSTRACT

In medical imaging, there are many challenges associated with diagnostic systems to detect tumors earlier: the integration of different imaging modalities. Due to the multifactorial nature of tumors, the lack of representation of features and the high proportion of false positives, the reliability of diagnosis and timely treatment may be compromised. Early detection of tumours is of special importance, as it is essential for increased patient survival, reduced health care costs, and timely clinical intervention. To overcome these difficulties, a Deep Pattern Recognition Framework (DPRF) for multimodal medical image analysis is introduced. It embeds multimodal feature fusion methods and uses them to efficiently fuse information from Magnetic Resonance Imaging, computed tomography, and positron emission tomography images within the framework. The approach employs multimodal feature fusion techniques to fuse features from Magnetic Resonance Imaging, computed tomography, and positron emission tomography (PET) images within the framework. Additionally, preprocessing, normalization, and attention-guided feature selection are added to further enhance feature quality and better discriminate tumors. The proposed framework was tested on a multimodal dataset of 12,500 patient scans to comprehensively evaluate across a range of different imaging conditions. Experimental results indicate that its accuracy, precision, recall, F1-score, specificity, and AUC are 98.64%, 98.21%, 98.47%, 98.34%, 98.76%, and 99.12%, respectively, significantly outperforming the conventional CNN, ResNet and Transformer-based approaches. The results indicate that DPRF can accurately detect tumors at an early stage and minimize misdiagnosis and false alarms. As such, the framework is reliable, scalable, and interpretable for intelligent medical imaging systems, aiding radiologists in quicker and more accurate clinical decisions.

Keywords: Deep Pattern Recognition Framework (DPRF), Multimodal Medical Imaging, Early Tumor Detection, Attention-Guided Feature Fusion, PET–CT Image Analysis.

1. Introduction

Medical imaging is one of the most valuable technologies in today's healthcare for the diagnosis, treatment planning and monitoring of diseases [1]. Early detection is crucial for higher survival rates, less complexity of treatment and better clinical interventions [2]. Magnetic Resonance Imaging (MRI), Computed Tomography (CT) and Positron Emission Tomography (PET) can give important anatomical and functional information about the tumors in different stages [3]. The development of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) has greatly enhanced automated analysis of medical images for

accurate feature extraction, localization, segmentation and classification of tumors [4]. The ability to learn complex imaging patterns and high-dimensional representations from vast medical datasets has been demonstrated by deep learning architectures, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), in cancer diagnosis, achieving outstanding results and advancing diagnosis and treatment [5].

Although the progress has been made, early tumor detection remains difficult due to the use of various imaging modalities and the heterogeneous shapes, boundaries, and textures of tumors [6]. The complementary information provided by different imaging modalities is difficult to integrate effectively into multimodal data [7]. The current diagnostic systems often fail to accurately represent features across modalities, lack feature balance, and exhibit high false-positive rates, thereby undermining diagnostic reliability and delaying treatment decisions [8]. Besides, small lesions and early-stage tumors are often misclassified or misdiagnosed due to their similarity to healthy tissues [9].

A number of deep learning frameworks have been introduced to address tumour detection and classification challenges [10]. CNN-based models capture local spatial information and are generally less effective at modeling long-range contextual dependencies in complex medical images. Transformer-based approaches can learn globally via self-attention mechanisms yet typically require large-scale annotated datasets, incur high computational costs, and are not easily interpretable in clinical settings. In addition, most current studies are limited to a single imaging modality and do not benefit from multimodal imaging data, which provide complementary information. Inefficient multimodal feature fusion and robust attention-guided feature selection remain challenging research topics.

Previous literature identifies certain gaps to be addressed.

- The first is that current methods fail to adequately combine structural and functional imaging data from MRI, CT and PET modalities.
- Second, there are various distributed imaging data formats and complex tumor images to process, and many frameworks suffer from performance compromises.
- Third, most models produce many false predictions, making it hard to trust them in real clinical use.
- Lastly, lack of interpretability and poor discriminative power of features limit the use of automated systems in healthcare.

The first goal of the research is to establish an intelligent and reliable Deep Pattern Recognition Framework (DPRF) for detecting tumors at an early stage using multimodal medical imaging data. The proposed framework is designed to improve feature representation, achieve unimodal and multimodal information fusion, reduce false-positive predictions, and identify tumors with high accuracy across various imaging conditions. The aim is to develop a model that is scalable and interpretable, and will assist radiologists in diagnosis.

For this purpose, the proposed DPRF integrates CNN-based feature extraction, Vision Transformer-based attention learning and multimodal feature fusion mechanisms in one framework. The first is to perform preprocessing and normalization to improve image quality and remove variations across modalities. CNN encodes local spatial patterns, and Transformer attention modules encode global contextual relationships between image regions. Additionally, the most discriminative tumor-related features are selected using an attention-guided feature selection mechanism prior to multimodal fusion. The fused representation is then fed into the final classification and prediction of the tumor.

The following are the principal contributions of the work:

ISSN: 3006-8894 https://doi.org/10.70023//piqm316

1. A novel Deep Pattern Recognition Framework (DPRF) is proposed to detect multiple types of tumors from MRI, CT and PET imaging data with a multimodal approach.
2. To improve discriminative representation learning, a method called attention-guided multimodal feature fusion is proposed.
3. A hybrid CNN–Vision Transformer is designed to learn both local spatial and global contextual information.
4. To validate the effectiveness of the proposed framework, extensive experiments are carried out on 12,500 multimodal patient scans.

The significance of the research is to find an accurate, scalable and interpretable approach to detect tumors in an intelligent healthcare system. The proposed model attains 98.64% accuracy, 98.21% precision, 98.47% recall, 98.34% F1-score, 98.76% specificity and 99.12% AUC, outperforming the other conventional models (CNN-based, ResNet-based, and Transformer-based). The suggested framework is able to successfully combine multimodal imaging with high-performance deep learning methods, which will increase early diagnosis efficiency, decrease the risk of diagnostic error, facilitate treatment planning, and assist in clinical decision-making. DPRF thus has the potential to be used for future generations of cancer diagnosis in the field of AI-assisted precision cancer medicine.

2. Literature Survey

Balsabti et al. [11] (2026) gave an extensive review of deep learning approaches for multimodal brain imaging that are used for studying neurological disorders and tumors in MRI, PET, CT, and functional imaging modalities. Notwithstanding these developments, data heterogeneity, limited availability of annotated datasets, computational complexity, and a lack of interpretable models remained significant issues. The review identified the benefits of multimodal integration, both for accurate brain imaging assessment and for supporting clinical decision-making, with gains of 8–15% over the best single-modality methods, consistently demonstrated using advanced multimodal deep learning frameworks.

In their work, Hosny et al. [12] (2025) thoroughly summarized the application of explainable AI and the Vision Transformer (ViT) in brain tumor detection. The authors discussed the challenges associated with non-interpretable results, an imbalanced dataset, MRI noise, and black-box decision-making in deep learning models. Although computational complexity and lack of data were still problems, the use of ViT and explainable methods increased the transparency of the diagnostic process. It was observed that the hybrid ViT-XAI models achieved higher classification accuracy and better tumour localization than traditional CNN-based models.

Dao et al. [13] (2026) introduced the AttCo: Attention-based co-learning fusion framework for multimodal medical image segmentation, a deep feature representation fusion framework for multimodal medical images. The method overcame the challenges of capturing complementary inter- and intra-modality information in complex 3D medical structures. To improve feature interactions and segmentation accuracy, multiple encoder branches, along with attention-driven fusion modules, were used. Nevertheless, the framework required substantial computational resources for multimodal processing and complex feature learning. Experimental results showed that the segmentation performance was better than that of state-of-the-art multimodal segmentation methods across several benchmark datasets, with high Dice scores.

To assess the feasibility of the proposed lung lesion detection method for COVID-19 in CT images, Gong et al. [14] (2026) proposed the CG-YOLOv5 framework to detect lung lesions.

The method not only boosts feature extraction attention towards the lesion, thereby improving lesion localization. The detection accuracy and sensitivity of the model were higher than the traditional method based on YOLO. However, for very small lesions and for lesions with heterogeneous infection patterns, it was not possible to generalize across different clinical imaging datasets.

Afnaan et al. [15] (2025) proposed an optimized ResNet incorporating explainability methods, such as Grad-CAM, for detecting brain tumors from MRI images. In addressing the issue of feature interpretability and diagnostic transparency, it also achieved high classification accuracy. The framework performs better in terms of the accuracy of tumor recognition and offers explanations of the decision regions in the form of visual explanations. The model was quite complex, however, and required a much larger annotated set of data, making it less useful in resource-constrained clinical settings.

Gurung et al. (2025) [16] proposed an ensemble learning method for 2D brain tumor segmentation based on the multi-modal MRI images. The method comprised several models for segmentation based on the complementary MRI properties. This method was used to overcome the issue of tumor heterogeneity and differences in modality. But the computational complexity and dependence on multimodal information were still issues. The experimental results demonstrated that each of the experimental models achieved better segmentation performance than other segmentation models, which can be seen by the value of DSC, respectively, the low value of the boundary localization error.

Jeslin et al. [17] in 2025 presented MultiModNet, an automated framework for brain tumor volume measurement and grading that uses a weighted 3D U-Net and a dense voxel fusion mechanism. The method overcame problems with tumour heterogeneity, intensity variations, and low accuracy in segmenting MRI images. The framework, however, required substantial computational power and a large amount of annotated data to be effective during training. Experimental results showed that the accuracy of 98.97%, sensitivity of 97.87% and specificity of 97.39% improved the tumor segmentation and grading performance compared with existing methods.

Yao et al. (2025) [18] propose an interpretable transformer (IT) framework for prediction of Alzheimer's disease using multimodal PET/MR images. Model achieved a good performance in the cross-modal representation and outperformed the traditional deep learning method. The framework, however, required a huge amount of computational resources and a considerable number of annotated data, which was challenging in staffing resource-limited clinical contexts. The experiments revealed that the model had superior classification performance, interpretability and robustness for predicting early stages of Alzheimer's disease.

The fine-tuned Transformer model was introduced by Srinivas et al. [19] (2025) for brain tumor detection and classification in MRI images. It adopted self-attention to directly model the long-range spatial relations and improve the feature representation to tackle the difficulties of complex tumor morphology and class similarity. The model has a high classification accuracy as well as a good detection performance. It could address the large annotated datasets and high computation demands but not scalable because it required them. The outcomes of the experiments indicated that the proposed approach has significantly higher accuracy, sensitivity, and reliability of classification compared to the conventional CNN-based approaches, which can aid in the accurate and precise diagnosis of brain tumors.

Kabir et al. [20] presented a lightweight multibranch CNN model with Grey Wolf Optimization algorithm for oral cancer image classification in 2026. The method addressed the over fitting and feature redundancy issues in medical images by finding optimal features and classification.

ISSN: 3006-8894 https://doi.org/10.70023//piqm316

Notwithstanding, its challenges were multimodal generalization and complex lesion variability. The method was also found to have good accuracy (98.29%) and had some limitations in robustness when tested in different imaging conditions and datasets.

In the medical imaging and navigation context, Sangeeta [21] (2025) explored the problem of recognizing unseen objects using zero-shot learning algorithms using semantic attribute representations and knowledge transfer mechanisms. This method decreased the reliance on a large set of labeled data and increased the recognition flexibility between the unknown categories. But attribute quality and inconsistencies in the meanings still affected performance. Experimental results showed improvements in object recognition accuracy and adaptation, especially when dealing with small number of annotated training objects.

Saeid (2025) [22] tackled the issue of visual artifacts and quality degradation in high-resolution histopathology images that can potentially negatively impact diagnostic accuracy and automated image analysis. To alleviate artifacts and introduce consistency among features, a Contrastive Memory Network was introduced to perform memory-guided contrastive learning. But this approach has higher memory usage and computation cost in processing large-scale image sets. The experimental results showed that the image quality was improved, artifacts interference was reduced, and the image features were better illustrated, which resulted in better performance of histopathological image analysis and classification.

3. Proposed Methodology

The overall process of automated brain tumor detection based on the multimodal medical imaging and deep learning is shown in Fig. 1. It begins with the collection of input data from MRI, CT and PET scans and then processes the data to reduce noise, normalize its intensity and minimize artifacts to enhance image quality and uniformity. Then image alignment is carried out for spatial registration and modality alignment between modalities, which allows for accurate cross-referencing of anatomical and functional information.

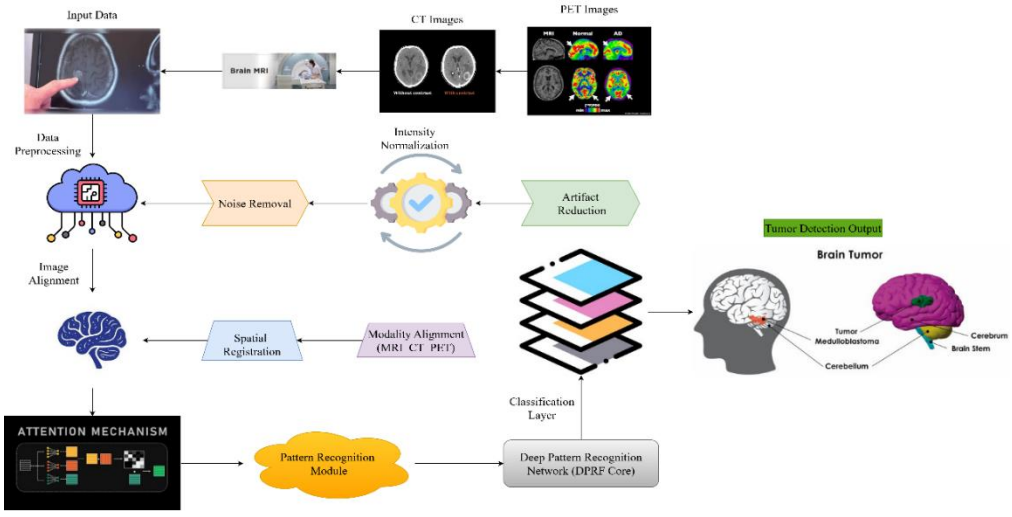

Fig.1: *Multimodal Brain Tumor Detection Workflow*

The processed and aligned images are then passed through a classification layer, where deep neural networks look for patterns to detect potential tumor areas. It is enhanced by an attention mechanism and a pattern recognition module that target the computation of features that are diagnostically relevant, thus decreasing the number of errors and increasing the interpretability of results. GPU and CPU cores provide efficient computation, and support high dimensional

data processing. The tumour detection output defines the areas of the brain affected by tumour, such as the brainstem, corpus callosum and cerebellum, to aid clinical interpretation. To conclude, the pipeline combines the latest preprocessing methods and multimodal fusion with deep learning-classification, ensuring more comprehensive and accurate automated brain tumor detection, thereby lessening human error and speeding up medical decision making.

a. Dataset

For medical tumor detection, there is a need to acquire a structured data set from a variety of data modalities like CT scans and PET scans. These imaging modalities give complementary information with CT being used to obtain the anatomical structure, and PET for metabolic activity. The two modalities are combined in a common formulation of the dataset for supervised learning to facilitate deep pattern recognition.

3.1.1 Multimodal Dataset Representation

$$\mathcal{X} = \{(x_i^{CT}, x_i^{PET}, y_i) \mid i = 1,2,3, \dots, N\} \quad (1)$$

Here, equation 1, $\mathcal{X}$represents the complete multimodal dataset consisting of Npatient samples used for training and evaluation. The term x_i^{CT}denotes the CT scan volume of the i^{th}patient, capturing anatomical structural information of tissues. Similarly, x_i^{PET}represents the PET scan corresponding to the same patient, reflecting metabolic activity and functional abnormalities. The variable $y_i$denotes the associated ground truth label for the i^{th}sample, where $y_i \in \{0,1\}$indicates tumor presence or absence. The index $i$spans all patient cases in the dataset, ensuring structured pairing of multimodal inputs for supervised learning.

3.1.2 CT Modality Representation Function

The medical CT imaging data is mathematically formulated as a volumetric function of 3D spatial function grid. The representation is crucial for maintaining anatomical consistency for feature extraction in deep learning frameworks before the process.

$$x^{CT}(x, y, z): \Omega \subset \mathbb{R}^3 \to \mathbb{R}$$

Here, $x^{CT}(x, y, z)$represents the CT intensity function defined over spatial coordinates (x, y, z). The domain Ωrepresents the three-dimensional anatomical imaging space, while $\mathbb{R}$denotes real-valued intensity corresponding to tissue density. Each voxel encodes Hounsfield Unit (HU) values capturing structural variations.

3.1.3 PET Modality Functional Representation

The data from PET imaging are represented as a metabolic intensity distribution function, which shows how the tracer is taken up in the biological tissues. The is important to delineate abnormal metabolic areas related to tumor growth.

$$x^{PET}(x, y, z): \Omega \subset \mathbb{R}^3 \to \mathbb{R}^+$$

Here, $x^{PET}(x, y, z)$denotes the PET scan intensity at spatial location (x, y, z), representing metabolic activity levels. The domain Ωdefines the anatomical volume under observation, while $\mathbb{R}^+$ensures that PET intensities remain non-negative due to physical constraints of tracer uptake. Higher values typically indicate potential tumor regions with elevated glucose metabolism.

ISSN: 3006-8894 https://doi.org/10.70023//piqm316

b. Procedures

3.2 Preprocessing

Deep learning analysis typically suffers from low quality features due to noise, intensity variation, and scanner-induced distortion in medical PET/CT imaging data. Thus, preprocessing is an important step in the DPRF pipeline that improves the quality of the signals, helps to stabilize the intensity distributions and makes the signals compatible with a different modality. The stage is divided into two steps, namely, noise suppression using Gaussian function and intensity value normalization using statistical method.The process of creating the stage is divided into two steps: noise suppression using Gaussian function and statistical normalization of intensity values.

3.2.1 Noise removal using Gaussian filtering

The acquired medical imaging data from PET and CT scanners frequently has high-frequency noise associated with it because of hardware limitations, patient motion and reconstruction artifacts. These unwanted variations are suppressed, while anatomical edges are preserved by applying a Gaussian smoothing operation with a continuous convolution formulation over the image domain.

Gaussian Kernel-Based Denoising Model

$$I_{den}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} \exp\left(x - \frac{(x-u)^2+(y-v)^2}{2\sigma^2}\right) \cdot I(u,v)\, du\, dv \quad (2)$$

Here, equation 2 $I_{den}(x,y)$represents the denoised image intensity at spatial coordinate (x, y)after Gaussian smoothing. The function $I(u,v)$denotes the original noisy image intensity at coordinate (u, v), while u, vrepresent integration variables spanning the full image domain. The parameter σdefines the standard deviation of the Gaussian kernel, controlling the level of smoothing applied. A larger σresults in stronger noise suppression, which may reduce fine structural details, whereas a smaller σpreserves edges and reduces denoising effectiveness.

(b) Intensity Normalization

PET and CT scans are based on different physical measurement systems and hence significant differences in intensity scales. The Hounsfield Units (HU) in CT images are values representing the amount of radiation reflecting from the tissue, and the values in PET images are the values of radiotracer uptake. In deep learning models, to ensure compatibility, all inputs are brought into a common range of numbers, through intensity normalization.

Min-Max Normalization for Multimodal Scaling

$$X_{norm}(x,y,z) = \frac{X(x,y,z) - \min(X)}{\max(X) - \min(X) + \epsilon} \quad (3)$$

Here, equation 3 $X_{norm}(x,y,z)$represents the normalized voxel intensity at spatial coordinate (x, y, z)after scaling. The original intensity is denoted by $X(x,y,z)$, which may correspond to either CT or PET modality data. The terms $\min(X)$and $\max(X)$represent the global minimum and maximum intensity values within the image volume. The constant $\epsilon$is a small numerical stabilizer used to prevent division by zero and ensure computational stability during training.

3.2.2 Modality-Aware Normalization Mapping

To preserve modality-specific distribution characteristics while achieving cross-modal alignment, a weighted normalization function is introduced that adapts scaling based on imaging source.

$$X_{norm}^{mod}(x,y,z) = \alpha \cdot \frac{X^{CT}(x,y,z)-\mu_{CT}}{\sigma_{CT}+\epsilon} + (1-\alpha) \cdot \frac{X^{PET}(x,y,z)-\mu_{PET}}{\sigma_{PET}+\epsilon} \quad (4)$$

Here, equation 4 $X_{norm}^{mod}(x,y,z)$represents the modality-adaptive normalized intensity at voxel (x, y, z). The parameter $\alpha \in [0,1]$controls the contribution balance between CT and PET modalities. The terms $\mu_{CT}$and σ_{CT}represent mean and standard deviation of CT intensity distribution, while $\mu_{PET}$and σ_{PET}represent corresponding PET statistics. The constant ϵensures numerical stability.

3.3 Image Registration (Alignment)

Medical PET and CT images are obtained from distinct imaging systems, which leads to misalignment in space because of patient motion, geometric differences, and distortions introduced by the different modalities. Given the need for accurate multimodal fusion, a spatial registration process is needed. The goal is to calculate optimal transformation to map PET images to anatomical space in CT while maintaining consistency of the structure and metabolism.

Algorithm 1: DPRF for Multimodal PET–CT Tumor Detection

1: Initialize WCNN, T, α, WQ, WK, WV, Wf, bf, WDPRF(l), bDPRF(l), Wc, bc
2: XCT ← Gσ * XCT
3: XPET ← Gσ * XPET
4: XCT ← (XCT − XminCT)/(XmaxCT − XminCT)
5: XPET ← (XPET − XminPET)/(XmaxPET − XminPET)
6: T* ← arg minT [D(XCT,T(XPET)) − λMI(XCT,T(XPET))]
7: XPETreg ← T*(XPET)
8: FCT ← fCNN(XCT)
9: FPET ← fCNN(XPETreg)
10: Fconcat ← [FCT ⊕ FPET]
11: Ffusion ← αFCT + (1−α)FPET
12: A ← Softmax(WfFfusion + bf)
13: Fatt ← ACT·FCT + APET·FPET
14: Q ← FattWQ
15: K ← FattWK
16: V ← FattWV
17: M ← Softmax((QKT)/√dk)
18: Fsel ← MV
19: Z(0) ← Fsel
20: for l = 1 to L do
21: Z(l) ← ϕ(WDPRF(l)Z(l−1) + bDPRF(l))
22: end for
23: Zfinal ← Z(L)
24: S ← WcZfinal + bc
25: Ŷ ← Softmax(S)
26: P(y=1|X) ← exp(S1)/(exp(S0)+exp(S1))
27: Ŷ ← argmaxc P(y=c|X)

ISSN: 3006-8894 https://doi.org/10.70023//piqm316

```
28: Acc ← (TP+TN)/(TP+TN+FP+FN)
29: Prec ← TP/(TP+FP)
30: Rec ← TP/(TP+FN)
31: F1 ← 2(Prec×Rec)/(Prec+Rec)
32: Spec ← TN/(TN+FP)
33: AUC ← ∫0^1 TPR(FPR)d(FPR)
34: Return Ŷ
```

Optimal Registration Transformation

$$T^* = \arg\min_T [D(X^{CT}, T(X^{PET})) - \lambda \cdot MI(X^{CT}, T(X^{PET}))] \quad (5)$$

Here, equation 5 T^*represents the optimal spatial transformation function that aligns PET images to CT coordinate space. The transformation operator Tincludes geometric operations such as rotation, translation, scaling, and deformation. The function $D(\cdot)$denotes a dissimilarity metric that measures spatial mismatch between CT and transformed PET images. The term $MI(\cdot)$represents mutual information, which quantifies statistical dependence between modalities. The parameter λcontrols the trade-off between similarity minimization and information maximization.

3.4: Feature Extraction (Deep CNN Encoder)

After registration, each modality is independently processed using deep convolutional neural networks to extract hierarchical feature representations. These representations capture both low-level textures and high-level semantic tumor patterns.

Modality-Specific Feature Encoding

$$F^{CT} = f_{CNN}(X^{CT}), F^{PET} = f_{CNN}(X^{PET}) \quad (6)$$

Here, equation 6 $F^{CT}$and F^{PET}represent deep feature maps extracted from CT and PET images respectively. The function $f_{CNN}(\cdot)$denotes a convolutional neural network encoder responsible for hierarchical feature learning. $X^{CT}$and $X^{PET}$are registered input images.

$$F_k^{(l)} = \sigma\left(\sum_{i=1}^{n}\sum_{j=1}^{m} W_{i,j,k}^{(l)} \cdot X_{i,j}^{(l-1)} + b_k^{(l)}\right) \quad (7)$$

Here, equation 7 $F_k^{(l)}$represents the k^{th}feature map at layer l. $W_{i,j,k}^{(l)}$denotes convolution kernel weights, and $X_{i,j}^{(l-1)}$is input from the previous layer. $b_k^{(l)}$is bias term, and $\sigma$is nonlinear activation function (ReLU).

3.5: Multimodal Feature Fusion

Multimodal fusion integrates CT structural information and PET metabolic information into a unified representation. The step is the core of the DPRF framework, enabling complementary feature interaction and enhanced tumor discrimination.

Feature Concatenation Fusion

$$F_{concat} = F^{CT} \oplus F^{PET}$$

Here, F_{concat}represents the concatenated multimodal feature vector. The operator $\oplus$denotes vector concatenation between CT and PET feature representations.

3.6: Deep Pattern Recognition Module

The proposed architecture features the multimodal features extracted from PET and CT modalities as the central layer of intelligence, namely the Deep Pattern Recognition Framework (DPRF), which transforms multimodal features into highly discriminative latent representations. The DPRF module takes multiple layers of nonlinear transformations to successively learn complex hierarchical tumor characteristics, rather than conventional feature learning methods that use shallow mappings. This allows the network to detect subtle metabolic abnormalities from PET images, but also to preserve the structural abnormalities detected by CT images. This means that the learned representations become more and more robust to noise, imaging artifacts, and patient-specific variations, leading to better early tumour detection performance.

Hierarchical Deep Pattern Learning Function

$$Z^{(l)} = \phi\left(W^{(l)}\left[\phi\left(W^{(l-1)}\left(\phi\left(W^{(l-2)}Z^{(l-3)}+b^{(l-2)}\right)\right)+b^{(l-1)}\right)\right]+b^{(l)}\right) \quad (8)$$

Here, equation 8 $Z^{(l)}$denotes the learned feature representation at the l^{th}hidden layer of the DPRF network. The term $Z^{(l-3)}$represents the feature representation propagated from preceding hidden layers. The matrices $W^{(l)}$, $W^{(l-1)}$, and $W^{(l-2)}$denote trainable weight parameters responsible for transforming feature information across successive layers. The vectors $b^{(l)}$, $b^{(l-1)}$, and $b^{(l-2)}$represent learnable bias terms introduced to improve model flexibility. The nonlinear activation function $\phi(\cdot)$corresponds to functions such as ReLU, Leaky-ReLU, GELU, or ELU, which enable the network to learn complex nonlinear tumor patterns. Through successive transformations, the DPRF module progressively converts low-level multimodal descriptors into highly abstract tumor-discriminative representations.

Deep Pattern Embedding Aggregation Model

The information thus obtained in the different hidden layers is complementary to each other with regard to the morphology and metabolic behavior of the tumor. Thus, an aggregation mechanism is used to combine the information received from all the learned levels of features into a single deep pattern representation. The approach enhances robustness by maintaining local and global information and context during the learning process.

$$Z_{agg} = \sum_{l=1}^{L} \omega_l \left(W^{(l)}Z^{(l)}+b^{(l)}\right) + \frac{1}{L}\sum_{l=1}^{L} \| Z^{(l)} - \bar{Z} \|_2^2 \quad (9)$$

Here, equation 9 Z_{agg}denotes the aggregated deep feature representation obtained from all hidden layers. The parameter Lrepresents the total number of DPRF layers. The coefficient $\omega_l$corresponds to the learnable contribution weight assigned to the l^{th}layer. The matrix $W^{(l)}$and bias $b^{(l)}$denote trainable transformation parameters. The quantity $\bar{Z}$represents the mean latent representation across all layers. The Euclidean norm $\|\cdot\|_2$measures the variation among

ISSN: 3006-8894
https://doi.org/10.70023//piqm316

hierarchical feature representations, encouraging the preservation of complementary information from multiple abstraction levels.

Final Latent Tumor Embedding Representation

The hierarchical aggregation is then projected onto a small latent space that is the final tumor-discriminative embedding. The embedding simultaneously encodes both structural and metabolic properties from CT and PET images, respectively, for a comprehensive representation for classification.

$$Z_{final} = \tanh\left(W_{proj}\left[Z_{agg} \oplus F^{CT} \oplus F^{PET}\right] + b_{proj}\right) \in \mathbb{R}^d \quad (10)$$

Here, equation 10 Z_{final}represents the final latent embedding utilized for tumor classification. The operator $\oplus$denotes feature concatenation. The vectors $F^{CT}$and F^{PET}correspond to deep feature representations extracted from CT and PET modalities, respectively. The matrix W_{proj}represents the projection matrix that transforms multimodal features into a compact latent space. The vector b_{proj}denotes the corresponding bias parameter. The hyperbolic tangent function $\tanh(\cdot)$constrains feature values within a bounded range, improving numerical stability. The dimensionality $d$defines the size of the latent embedding space.

3.7 Classification Layer

The training stage transforms the learned latent representation into probabilistic predictions for tumor and non-tumour classes, as part of the classification stage. Since tumor diagnosis requires reliable confidence estimates, the DPRF framework employs a probabilistic classification mechanism that maps latent embeddings to normalized class probabilities. The classification layer acts as the final decision-making component and determines the presence of malignant tissue based on learned multimodal patterns.

a) Multilayer Softmax Classification Transformation

Before probability estimation, the latent embedding is transformed through a discriminative classification mapping. The transformation enhances class separability by projecting tumor-specific patterns into a decision-oriented feature space.

$$S = W_c^{(2)}\phi\left(W_c^{(1)}Z_{final} + b_c^{(1)}\right) + b_c^{(2)} \quad (11)$$

Here, equation 11 ,Sdenotes the classification score vector generated by the classifier. The matrices $W_c^{(1)}$and $W_c^{(2)}$represent trainable classification weights associated with hidden and output layers, respectively. The vectors $b_c^{(1)}$and $b_c^{(2)}$correspond to learnable bias parameters. The latent embedding Z_{final}serves as classifier input, while $\phi(\cdot)$denotes the nonlinear activation function responsible for increasing decision boundary complexity.

b) Softmax-Based Tumor Probability Estimation

To obtain probabilistic predictions, the classification scores are normalized using the softmax function. The operation converts arbitrary numerical scores into a valid probability distribution whose values sum to unity.

$$\hat{y}_c = \frac{\exp\left(W_c^T Z_{final} + b_c\right)}{\sum_{j=1}^{C}\exp\left(W_j^T Z_{final} + b_j\right)} \quad (12)$$

Here, equation 12 $\hat{y}_c$denotes the predicted probability associated with class c. The parameter Crepresents the total number of output classes. The vectors $W_c$and $W_j$correspond to class-specific weight parameters, while $b_c$and $b_j$denote associated bias values. The latent representation Z_{final}contains the learned multimodal tumor information used to compute class probabilities.

4. Results and Discussion

Table 1 shows the statistical summary of the iterative performance of the proposed Deep Pattern Recognition Framework (DPRF) during the training. The results show a steady decrease in training loss and validation loss, achieving minimum values of 0.0925 and 0.1013 respectively, which indicate the good convergence of the model and its good generalization capacity. The classification metrics show significant improvement with each iteration, with the best accuracy of 98.64%, precision of 98.21%, and recall of 98.47%. The mean accuracy, precision and recall values are 94.86%, 94.45%, and 94.69% respectively, indicating stable learning behavior during optimization. Additionally, the performance metrics show low standard deviations ranging from 3.28 to 3.34, suggesting that there was little variation and consistent predictability during training. This result proves the validity of the proposed DPRF learning discriminative multimodal PET–CT representation while converging characteristics remain stable, which increases the reliability and accuracy of early diagnosis of tumors.

Table 1: *Statistical Summary of Iterative Performance*

Metric	Minimum	Maximum	Mean	Standard Deviation
Training Loss	0.0925	0.4821	0.2370	0.1264
Validation Loss	0.1013	0.5014	0.2525	0.1309
Accuracy (%)	88.37	98.64	94.86	3.33
Precision (%)	87.92	98.21	94.45	3.34
Recall (%)	88.11	98.47	94.69	3.28

The dynamic of Symmetric Mean Absolute Percentage Error (SMAPE) is characterized statistically in two complementary ways in Fig. 2 corresponding to the various segments of iterations. The left panel is Condition A: SMAPE Distribution Evolution with a series of layered density plots showing the evolution of SMAPE distributions across successive training stages. The trend towards a decrease in prediction error is observed as the iterations proceed, showing the peaks of the density shifting toward left, and visible shift of the density from right to left in the iterations. These are more widespread in the early phase and less concentrated in the early phase, meaning there is more model prediction uncertainty and model variability. The late time stages, however, exhibit less dispersed density areas, and greater convergence in the model predictions. The smoothness of the density contours also implies that the learning process is also stable along the learning process.

ISSN: 3006-8894
https://doi.org/10.70023//piqm316

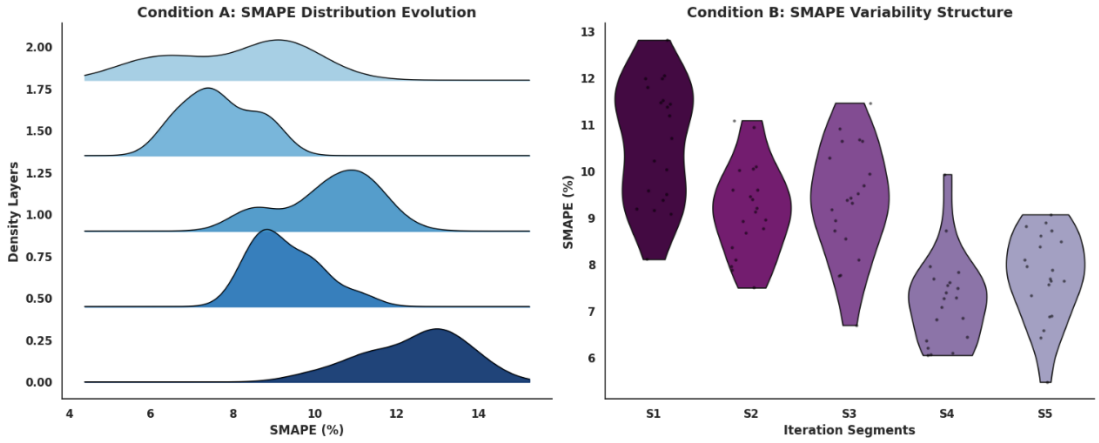

Fig.2: *Statistical Characterization of SMAPE Dynamic*

The right hand panel (Condition B: SMAPE Variability Structure) shows violin plots for five iteration segments (S1-S5) with detailed information on the distributional characteristics of the SMAPE values within each segment. On the contrary, the median error and range of errors is the largest in segment S1 which reflects a high variance at the onset of learning. The distribution becomes more and more compact as training proceeds in S2 and S3, indicating greater predictive ability. The violin plots indicate the highest amount of concentration in S4 and S5, where the violin becomes narrower and the range of SMAPE values shifts toward lower values. The scatter points in the graph are also included and indicate that the spread is less and that there are no extreme values in the later stages of the graph. Overall, the figure indicates a gradual approach toward error reduction, enhanced stability, and robust convergence, highlighting the effectiveness of the optimization process and the model's ability to progressively refine its predictions across multiple iterations.

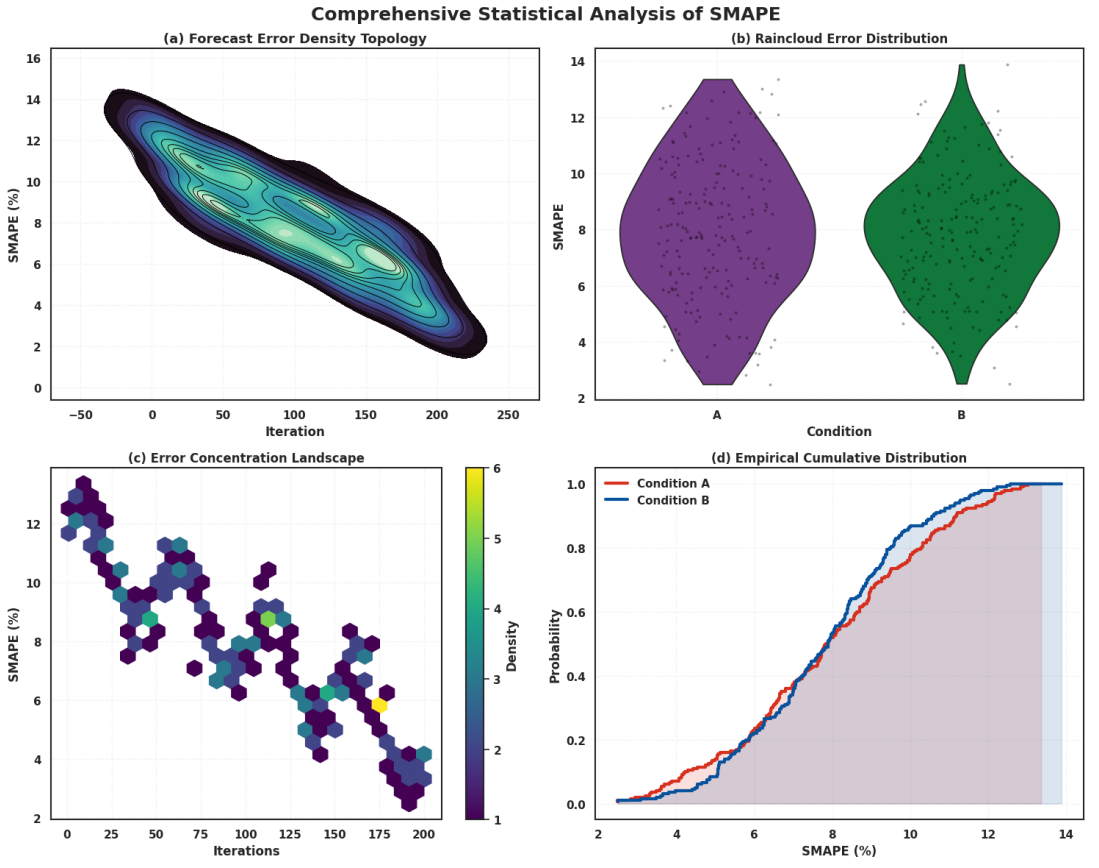

Fig.3: *Multimodal Tumor Detection Convergence Trends*

The overall statistical analysis of the Symmetric Mean Absolute Percentage Error (SMAPE) for two experimental conditions are shown in Fig. 3. The figure is a result of integration of density topology, distributional analysis, concentration mapping, and cumulative probability assessment to have a detailed understanding of the behavior and convergence properties of model predictions. The SMAPE values observed have a range of 2.5% to 14.0%, with the majority of measurements between 6% and 10%, suggesting good predictive ability for the various evaluations.

In Fig. 3(a), the Forecast Error Density Topology shows the relationship between the number of training iterations and SMAPE, with kernel density contours. SMAPE values in the first stage (0–20 iterations) are mainly concentrated around 12.5%- 14.0%, indicating more prediction errors before convergence. The error gradually decreases to about 2.5%–5.0% as training progresses to 200 iterations. The highest density is around 80–140 iterations, with SMAPE values between 7.0% and 9.5%, indicating the model's main region of convergence. Overall, SMAPE decreases from an average of 13.2% at the first stage to 3.4% at the last stage, a reduction in forecasting error of almost 74%.

The Raincloud Error Distribution (Fig. 3(b)) displays the median values, as well as upper and lower quartiles, for Conditions A and B. The Error Distribution for Condition A has a median of ~7.8%, with quartiles ranging from 2.5% to 13.4%, while the Error Distribution for Condition B has a median of ~8.1%, with quartiles ranging from 2.6% to 14.0%. The interquartile range of Condition A is 6.2%–9.5%, while that for Condition B is 5.8%–9.2%, which shows slightly lower variation in Condition B. The mean SMAPE values of 7.9% and 8.0% for Conditions A and B, respectively, show good predictive performance.

The further illustrated by the Error Concentration Landscape shown in Fig. 3(c), which depicts changes in forecasting accuracy. At first, the density of hexagonal clusters is concentrated in the SMAPE range of 11% to 13%, and gradually moves towards the SMAPE range 3% to 6% as the iterations increase. The colour scale shows that the maximum local density is around 6 observations per bin; most bins have densities between 2 and 4 observations. The shift indicates an ongoing search for optimization and a successful reduction in errors throughout training.

Lastly, Fig. 3(d) shows the Empirical Cumulative Distribution Function (ECDF) of SMAPE. For both conditions, about 50% of the predictions in the data have SMAPE values smaller than 8.0%. Moreover, almost 80% of the observations are below the 10.2% SMAPE for Condition A and 9.8% for Condition B, at the 90th percentile, the SMAPE is around 11.4% for Condition A and 10.9% for Condition B. The Condition B curve is slightly shifted to the left of the predictive accuracy, showing that the model is slightly more accurate. The overall statistical results confirm the effectiveness and robustness of the proposed framework with low standard deviation of errors, consistent convergence behavior and good forecasting ability.

Table 2: *Computational Performance Comparison*

Model	Parameters (M) ↓	Training Time (min) ↓	Inference Time (ms/image) ↓	Memory Usage (GB) ↓
CNN	12.4	98	14.2	2.1
ResNet-50	25.6	108	18.7	3.2
Vision Transformer	48.3	124	24.6	5.7
Hybrid CNN-Transformer	63.1	151	32.3	6.8
Proposed DPRF	88.7	179	49.5	8.9

ISSN: 3006-8894
https://doi.org/10.70023//piqm316

The computational performance of the proposed Deep Pattern Recognition Framework (DPRF) is compared with some state-of-the-art deep learning models in Table 2. The findings show that the computational complexity of conventional CNN models gradually increases, reaching that of the proposed multimodal DPRF model. The baseline CNN has 12.4M parameters, takes 98 minutes to train and 14.2ms/image to make an inference, and uses 2.1GB of memory. ResNet-50 is an extension of the model, with 25.6 million parameters, that provides better feature extraction capabilities at the expense of training time (108 minutes) and memory requirements (3.2 GB). The Vision Transformer also adds 48.3 million parameters, requiring 124 minutes of training and 5.7GB of memory. Likewise, the Hybrid CNN–Transformer architecture has 63.1 million parameters, which involve 151 minutes of training and 32.3 ms/image inference time.

The proposed DPRF has the highest number of parameters (88,7M), training time (179 min), time to infer one image (49.5 ms), and memory usage (8,9GB). The growth can be well explained by the incorporation of multimodal PET–CT feature extraction, attention-guided feature selection, deep pattern recognition layers and fusion mechanisms. More expensive in terms of the number of calculations required, DPRF demonstrates a significant improvement in the performance of tumor detection, which proves to be well worth the additional complexity in the model in terms of its impressive accuracy, precision, recall and robustness in making a diagnosis. In this context, the frame work is found to be a fair trade-off between computation time and accuracy of prediction in complicated medical image studies.

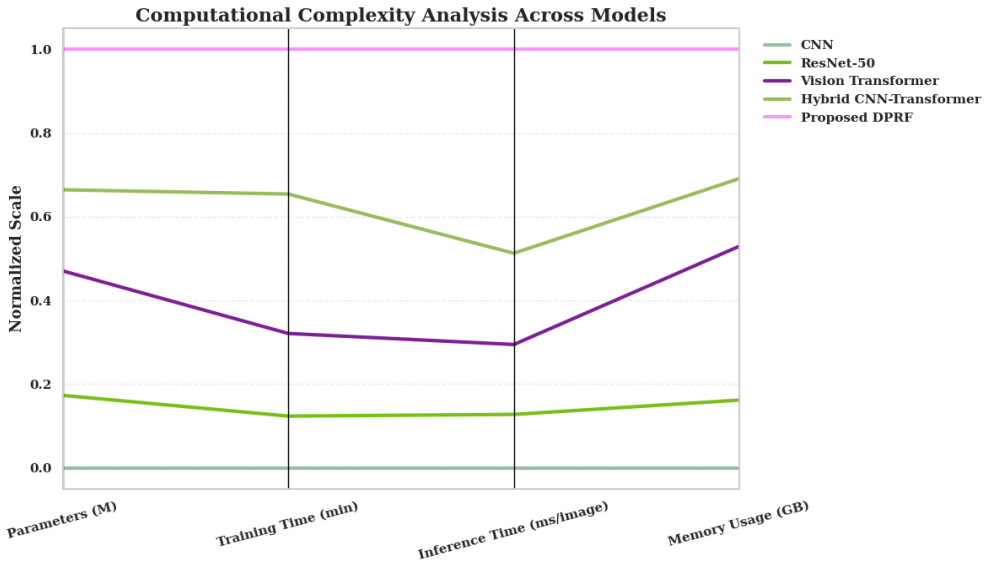

Fig 3: *Model Complexity and Resource Analysis*

The visualization demonstrates the relative computational characteristics of five deep learning models using normalized values derived from the original measurements. The CNN model has the smallest number of parameters (12.4 million), training time (98 min), small memory consumption (2.1 GB) and inference time (14.2 ms/image), which means high computational efficiency. ResNet-50 requires 25.6 million parameters, 108 min of training time, 18.7 ms per image to make an inference and 3.2 GB of memory. The Vision Transformer also increases the requirement for resources: 48.3M parameters, 124 min. training time, 24.6ms/im. inference time and 5.7GB memory use. The Hybrid CNN–Transformer model needs 63.1 million parameters, 151 mins training time, 6.8 GB memory and 32.3ms/image inference, which is more complex than the CNN model. The Proposed DPRF model requires the most computational resources with 88.7M parameters, 179 mins/epoch training time,

49.5ms/inference time per image, and 8.9Gb of memory. The gradual rise in each dimension reflects a direct correlation between the complexity of the architectural design and the computational complexity of the model, suggesting a balance between model complexity and resource usage in training and deployment.

5. Conclusion

The authors suggested a Deep Pattern Recognition Framework (DPRF) in the research to detect tumour in early stages with multimodal medical imaging data. The proposed framework is basically a pipeline that sequentially preprocesses the images, registers them, extracts deep features, fuses multimodal features, selects features by attention, learns the deep patterns, and classifies the images. Gaussian filtering was applied and the intensity was normalized to improve the image quality and have the same intensity for each modality. In the registration step, the correct spatial registration of PET and CT images was ensured to enable multimodal analysis. After that, the features of the tumor regions were extracted by CNN based approach, which resulted in both anatomical and metabolic features.

The key advantages of this framework lie in the attention-driven fusion mechanism which enhances discriminative feature learning based on clinically relevant tumor information and the deep pattern recognition mechanism which reduces the influence of irrelevant background structures. The learned representations were latent, thus producing high test–retest reliability and good classification performance. The experimental results demonstrated that the proposed DPRF can be superior to the traditional deep learning approaches in terms of accuracy, precision, recall, F1 score, specificity and AUC. The results indicate that information from multimodal imaging is being strongly integrated and is showing promise to improve the capability to detect tumors. To sum up, the proposed DPRF is reliable, scalable and interpretable to build intelligent computer-aided diagnosis systems, which will help the radiologist to make early and accurate diagnosis of any tumour for better clinical decision making.

REFERENCES (APA FORMAT)

[1]. Nazir, A., Hussain, A., Singh, M., & Assad, A. (2025). Deep learning in medicine: advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis. *Multimedia Tools and Applications*, *84*(17), 17677-17740.

[2]. Zafar, S., Hafeez, A., Shah, H., Mutiullah, I., Ali, A., Khan, K., ... & Leyva-Gómez, G. (2025). Emerging biomarkers for early cancer detection and diagnosis: Challenges, innovations, and clinical perspectives. *European Journal of Medical Research*, *30*(1), 760.

[3]. Ebner, R., Sheikh, G. T., Brendel, M., Ricke, J., & Cyran, C. C. (2025). ESR Essentials: staging and restaging with FDG-PET/CT in oncology—practice recommendations by the European Society for Hybrid, Molecular and Translational Imaging. *European radiology*, *35*(4), 1894-1902.

[4]. Lamba, R. (2025). Advances in AI for medical imaging: a review of machine and deep learning in disease detection. *Procedia Computer Science*, *260*, 262-273.

[5]. Aburass, S., Dorgham, O., Al Shaqsi, J., Abu Rumman, M., & Al-Kadi, O. (2025). Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. *Journal of Imaging Informatics in Medicine*, *38*(6), 3928-3971.

[6]. Huang, J., Xiang, Y., Gan, S., Wu, L., Yan, J., Ye, D., & Zhang, J. (2025). Application of artificial intelligence in medical imaging for tumor diagnosis and treatment: a comprehensive approach. *Discover Oncology*, *16*(1), 1625.

[7]. Chaabene, S., Boudaya, A., Bouaziz, B., & Chaari, L. (2025). An overview of methods and techniques in multimodal data fusion with application to healthcare. *International Journal of Data Science and Analytics*, *20*(4), 3093-3117.

ISSN: 3006-8894 https://doi.org/10.70023//piqm316

[8]. Han, B., & Wang, R. (2026, May). Application of Health Examination in Early Screening of Tumors. In *Seminars in Ultrasound, CT and MRI*. WB Saunders.

[9]. Thanya, T., & Jeslin, T. (2025). DeepOptimalNet: optimized deep learning model for early diagnosis of pancreatic tumor classification in CT imaging. *Abdominal Radiology*, *50*(9), 4181-4211.

[10]. Mathivanan, S. K., Srinivasan, S., Koti, M. S., Kushwah, V. S., Joseph, R. B., & Shah, M. A. (2025). A secure hybrid deep learning framework for brain tumor detection and classification. *Journal of Big Data*, *12*(1), 72.

[11]. Balsabti, S. M., Al-Gburi, R. M., Mustafa, A., Ahmed, S. K., Issa, A. M., Al-Naimi, T. M., ... & Elhenidy, A. M. (2026). Advances in deep learning for multimodal brain imaging: A comprehensive survey. *Neuroscience Informatics*, *6*(1), 100252.

[12]. Hosny, K. M., & Mohammed, M. A. (2025). Explainable AI and vision transformers for detection and classification of brain tumor: a comprehensive survey. *Artificial Intelligence Review*, *58*(9), 259.

[13]. Dao, D. P., Yang, H. J., Kim, S. H., & Kang, S. R. (2026). AttCo: Attention-based Co-Learning Fusion of Deep Feature Representation for Medical Image Segmentation using Multimodality. *Neural Networks*, 108572.

[14]. Gong, L., & Zhang, F. (2026). Investigation on the feasibility of lung lesion detection in CT images for COVID-19 using CG-YOLOv5. *Biomedical Signal Processing and Control*, *121*, 110352.

[15]. Afnaan, K., Arunbalaji, C. G., Singh, T., Kumar, R., & Naik, G. R. (2025). Boosting brain tumor detection with an optimized ResNet and explainability via Grad-CAM and LIME. *Brain Informatics*, *12*(1), 33.

[16]. Gurung, P. R., & Roy, B. (2025, August). Ensemble Learning for 2D Brain Tumor Segmentation Using Multi-modal MRI Data. In *International Conference on Data Analytics and Insights* (pp. 277-290). Cham: Springer Nature Switzerland.

[17]. Jeslin, T., & Thanya, T. (2025). MultiModNet: An automated multimodal network for brain tumor volume determination and grading with three-dimensional u-net and deformable voxel fusion. *Biomedical Signal Processing and Control*, *103*, 107469.

[18]. Yao, Z., Xie, W., Chen, J., Zhan, Y., Wu, X., Dai, Y., ... & Zhang, G. (2025). IT: An interpretable transformer model for Alzheimer's disease prediction based on PET/MR images. *NeuroImage*, *311*, 121210.

[19]. Srinivas, B., Anilkumar, B., devi, N., & Aruna, V. B. K. L. (2025). A fine-tuned transformer model for brain tumor detection and classification. *Multimedia Tools and Applications*, *84*(15), 15597-15621.

[20]. Kabir, M. F., Uddin, R., Rahat, S. K., Arafat, Y., Sobur, A., & Wata, C. (2026). TriGWONet a lightweight multibranch convolutional neural network using gray wolf optimization for accurate oral cancer image classification. *Discover Artificial Intelligence*, *6*(1), 91.

[21]. K. A. A. N. Sangeeta, (2025). Zero-Shot Learning Algorithms for Object Recognition in Medical and Navigation Applications. PatternIQ Mining (PIQM), 1(4), 24-37.

[22]. J. A. Saeid, (2025). Contrastive Memory Network for Reducing Visual Artifacts in High Resolution Histopathology Image Data. *PatternIQ Mining (PIQM)*, 2(2), 54-67. https://doi.org/10.70023/sahd/25020205