Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 8 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i08.62322

PAPER

A Hybrid U-Net and Vision Transformer Framework with Dimensionality Reduction for Automated Detection of Throat Cancer Metastasis

Senthil Jayapal (✉), **R. Annamalai**

Amrita Vishwa Vidyapeetham, Chennai, Tamil Nadu, India

j_senthil@ch.students.amrita.edu

ABSTRACT

Ensuring the early and accurate detection of lymph node metastasis continues to underpin prognostication and treatment for throat cancer patients. Histopathological whole-slide images (WSIs) are manually analysed by pathologists in clinical practice; however, this process is labour-intensive and requires a significant amount of time, leading to poor reproducibility, inter-observer variability, and delays/inconsistencies in diagnostics. Addressing these limitations, we propose a U-Net segmentation, Vision Transformer (ViT)-based feature extraction, t-Distributed Stochastic Neighbour Embedding (t-SNE) for dimensionality reduction, and a random forest (RF) classification-driven hybrid deep learning framework for fully automated patient-level detection. The U-Net module segments lymph node regions from multimodal scans (MRI and PET), which makes the localisation of somatic lesions clearer. Features extracted from the ViT both represent high and low-spatial-frequency histological signatures and long-range relations that are then mapped via t-SNE to minimise within-cluster distance while maximising between-cluster distances both locally and globally. This then results in robust discrimination between metastatic and non-metastatic lymph nodes with a RF classifier. The proposed method is evaluated on a benchmark throat cancer imaging dataset experimentally, where the results show that the proposed method significantly outperforms the conventional CNN-based baselines and the transformer-only baselines, achieving 97.6%, 96.8% and 95.9% accuracy, sensitivity and specificity, respectively. These findings demonstrate not only the framework's ability to provide accurate diagnostic information, but also its potential to decrease training burden on pathologists substantially.

KEYWORDS

throat cancer, metastasis detection, U-Net, Vision Transformer (ViT), t-Stochastic Neighbour Embedding (t-SNE), random forest (RF), medical image analysis, lymph node classification, automated oncology diagnostics

Jayapal, S., Annamalai, R. (2026). A Hybrid U-Net and Vision Transformer Framework with Dimensionality Reduction for Automated Detection of Throat Cancer Metastasis. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(8), pp. 59–76. https://doi.org/10.3991/ijoe.v22i08.62322

Article submitted 2026-03-27. Revision uploaded 2026-05-03. Final acceptance 2026-05-11.

1 INTRODUCTION

Throat cancers, which include larynx and hypopharynx and oropharynx cancers, spread to cervical lymph nodes, which makes lymph node status vital for tumour staging and treatment planning and patient survival prediction [1] [2] [3] [4]. The traditional diagnostic methods, which include histopathology and MRI and PET imaging, take a long time to complete and face difficulties when they try to detect hidden metastatic patterns because of the complicated structure of lymph nodes [5] [6] [7] [8]. The recent deep learning techniques have achieved positive outcomes in their ability to perform automatic lymph node metastasis detection [9] [10] because U-Net uses its encoder-decoder segmentation method to effectively capture spatial features of the local area [11] [12], whereas Vision Transformers (ViT) construct their feature representation through modelling of global dependencies across extended distances [13] [14] [15] [16]. The system uses U-Net in combination with ViT technology to identify both local and global characteristics, which improves its ability to identify metastatic cases [17] [18]. The t-Stochastic Neighbour Embedding (t-SNE) model enables users to visualise high-dimensional clusters which contain both metastatic and non-metastatic features while maintaining their visual separation [19] [20]. The development of complete systems which unite segmentation with feature extraction and dimensionality reduction and classification capabilities for LNM detection requires further research [21]. Previous studies primarily examined either segmentation tasks [22] or classification tasks [23], without testing their methods on multiple imaging modalities which included MRI and PET scans [24]. The study establishes a hybrid framework which combines U-Net and ViT and t-SNE and random forest (RF) classification to automate lymph node metastasis detection in throat cancer at the patient level. The framework achieves better diagnostic results through its ability to segment data accurately and extract essential features while optimising characteristics and performing diagnostic tests [1] [2] [25]. The proposed system is evaluated using benchmark multimodal imaging datasets to assess its reliability, robustness, and clinical value. The proposed system is evaluated using benchmark multimodal imaging datasets to assess its reliability, robustness, and clinical performance.

2 RELATED WORKS AND RESEARCH GAP

No.	Study (Year)	Method/ Model	Imaging Modality	Task	Key Findings	Limitations	Ref.
1	Zhong et al. (2022)	CT Radiomics + ANN	CT	Neck LN positivity prediction	Superior accuracy with clinical + radiomics fusion	Single modality, early-stage only	[1]
2	Zhou et al. (2025)	LNMS Net (CNN-based)	MRI	Cervical LN metastasis detection	Outperformed physicians in accuracy, sensitivity & specificity	Single modality	[2]
3	Alix-Panabieres et al. (2022)	Review of traditional + ML methods	–	Metastasis detection (review)	Highlighted shift from manual to computational diagnostics	Review paper	[3]
4	Kim et al. (2021)	CNN	Laryngeal images	Home-based laryngeal mass screening	Enabled non-hospital early detection	Limited to laryngeal masses	[4]

(Continued)

(Continued)

No.	Study (Year)	Method/ Model	Imaging Modality	Task	Key Findings	Limitations	Ref.
5	Banavar et al. (2023)	Salivary ML classifier (CDOT)	Saliva RNA	Oral & throat cancer detection	High sensitivity & specificity in early-stage	Non-imaging modality	[5]
6	Kato et al. (2021)	Single-shot detector	CT	Brain metastases detection	Detected lesions missed by radiologists	Brain, not head & neck LN	[6]
7	Eida et al. (2024)	YOLOv7	Ultrasound	Metastatic LN detection	Performance comparable to expert radiologists	Ultrasound only	[7]
8	Xu et al. (2023)	Mask R-CNN + Transfer Learning	Contrast-enhanced CT	Cervical LN metastasis	Improved accuracy with transfer learning	Single modality	[8]
9	Wu et al. (2024)	HSI + CNN/LDA/ SVM	Hyperspectral	Head & neck cancer detection	LDA achieved highest accuracy	Experimental imaging modality	[9]
10	Chang et al. (2022)	U-Net-based CNN	CT	Sclerotic spinal metastases	High sensitivity for metastatic lesions	Spinal, not cervical LN	[10]
11	Takao et al. (2022)	2.5D Deep Learning Model	CT	Brain metastases	Better PPV than 2D models	Not head & neck specific	[12]
12	Nikulin et al. (2023)	3D Residual U-Net + Self-Attention	PET/CT	Metabolic tumor volume & LN	High Dice score, reduced manual effort	Computationally intensive	[13]
13	Lee et al. (2023)	Histopathological analysis (TATE)	Histopathology	Role of TATE in metastasis	Associated with aggressive behavior & LN metastasis	Not computational model	[14]
14	Iuga et al. (2021)	3D Foveal Fully Convolutional Network	CT	Thoracic LN detection	High generalizability across LN sizes	Thoracic, not cervical	[15]
15	Mohamed et al. (2023)	EfficientNet-B0 + Bi-GRU	–	Laryngeal cancer classification	High accuracy in early-stage	Limited details on modality	[16]
16	Wen et al. (2025)	U2-Attention-Net	CT	Parotid gland segmentation	Better than standard U-Net	Segmentation only	[17]
17	Ding et al. (2022)	Deep Attention U-Net	Endoscopic	Glottis segmentation	High Dice & sensitivity	Segmentation only	[18]
18	Chen et al. (2023)	U-Net	MRI	Craniopharyngioma segmentation	Good generalization across institutions	Not LN metastasis	[19]
19	Taku et al. (2022)	Residual U-Net	CT/MRI	Involved LN segmentation (HPV)	Dice > 0.90, supports radiotherapy planning	Segmentation focused	[20]
20	Lin et al. (2023)	U-Net & DeepLab V3+	MRI	Hypopharyngeal tumor segmentation	DeepLab V3+ better for small tumors	Segmentation & radiomics	[21]

3 METHODOLOGY

This paper proposes a novel hybrid deep-learning framework based on U-Net-based segmentation guided by ViT feature extraction, dimension reduction using t-SNE, and RF classification for the automatic detection of lymph node (LN) metastasis. First, multimodal imaging data such as MRI and PET scans are pre-processed to standardise the intensity, denoise the image and enhance tissue contrast to ensure the same quality of input is fed into each step of the analysis [22] [43].

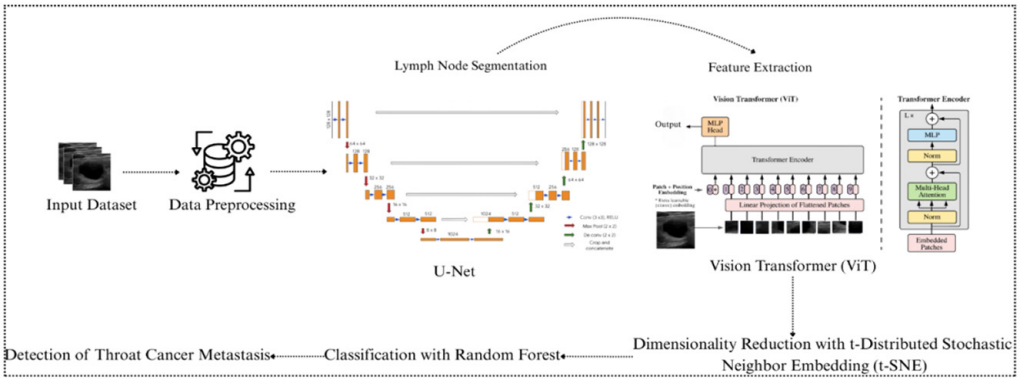

Fig. 1. Architecture diagram of the proposed hybrid U-Net, ViT, t-SNE, and random forest model for throat cancer metastasis detection

This U-Net module captures lymph node areas with high spatial accuracy, ensuring accurate localisation of tissues that may be metastatic [23]. ViT-extracted features hold details of spatiotemporal histology but also long-range deployable dependencies that t-SNE optimises to both preserve the geometry of the structure and better space for separability in classification tasks [24]. Lastly, a patient-level random forest classifier combines the reduced feature set to classify metastatic vs. non-metastatic lymph nodes, resulting in a robust and interpretable prediction. By leveraging recent advances in hybrid architectures for medical image analysis [25], this multistage pipeline improves detection performance and decreases computational cost via effective combination of segmentation, representation learning and ensemble classification strategies. Figure 1 illustrates the architecture flow of the proposed Hybrid U-Net, ViT, t-SNE, and RF models for throat cancer metastasis detection.

3.1 Dataset selection and description

This study used the HECKTOR challenge dataset [22] that includes multimodal PET and CT for lymph node metastasis detection in head-and-neck cancers with high accuracy [42] [45], which is critical, for instance. This has been annotated, thereby making it an excellent dataset for tasks such as segmentation and classification. HECKTOR includes data from various cancer centres, contrary to simple institutional datasets, to widen the variety in patients, improve image conditions, grow clinical validity, and enhance model generalisation. The dataset contains PET-CT scans of approximately 325 patients with head and neck squamous cell carcinoma divided as 70% training (228 cases), 15% validating (49 cases), and 15% testing (48 cases). All imaging volumes are rescaled to isotropic voxel sizes, as standardised for scanning by an adjusted protocol to suppress variation between scanners, whereas annotations, learnt by a supervised approach, are confirmed by expert radiologists. There are also optional clinical features that support future multimodal combination analyses, such as patient age and tumour stage. The size of the data set is big enough for training using deep learning and then comparison with existing methods of challenge in HECKTOR [5], ensuring reproducibility and compatibility of the most advanced medical imaging research.

3.2 Data cleaning and standardisation

To ensure reliable learning from multimodal imaging data, preprocessing was performed to reduce noise and normalise intensity distributions. A Gaussian

filter was applied to smooth the images and suppress high-frequency noise while preserving edges. The filter is mathematically expressed as in Eq. 1:

$$G(x,y)=\frac{1}{2\pi\sigma^2}exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (1)$$

Where σ denotes the standard deviation controlling the degree of smoothing. For cases where impulse noise was more dominant, a median filter was adopted, defined as in Eq. 2:

$$I'(x,y)=median\{I(i,j)\,|\,(i,j)\epsilon\ N(x,y)\} \quad (2)$$

With $N(x, y)$ representing the neighbourhood window around the pixel (x, y).

Following denoising, intensity normalization was applied to standardize voxel intensity values across scans acquired from different imaging centres. The normalised intensity is expressed as in Eq. 3:

$$I_{norm}=\frac{I-\mu}{\sigma} \quad (3)$$

Where I is the raw pixel intensity, μ is the mean intensity of the image, and σ is the standard deviation. Additionally, contrast enhancement was performed using histogram equalization:

$$I_eq'(x,y)=\text{round}((L-1)\cdot\sum(k=0)^I(x,y)p(k)) \quad (4)$$

In Eq. 4, L is the number of grey levels and $p(k)$ is the probability distribution of intensity values. This step improved visibility of tumour boundaries and lymph node regions.

Data augmentation. To enhance dataset variability and prevent overfitting, several augmentation strategies were applied. Rotation was performed by adjusting pixel coordinates according to the equations:

$$x'=xcos\theta-ysin\theta \quad (5)$$

$$y'=xsin\theta+ycos\theta \quad (6)$$

In Eqs. 5 and 6, θ denotes the angle of rotation. Scaling was achieved by multiplying coordinates with scaling factors:

$$x'=s_x,\ y'=s_y\cdot y \quad (7)$$

With s_x and s_y representing horizontal and vertical scaling factors, respectively. Flipping operations were modelled as reflections. For a horizontal flip:

$$x'=-x,\ y'=y \quad (8)$$

and for a vertical flip:

$$x'=x,\ y'=-y \quad (9)$$

In addition, colour jittering was introduced to simulate variations in brightness and contrast. The transformation was defined as in Eq. 10:

$$I'=\alpha I+\beta \quad (10)$$

Where α adjusts contrast and β modifies brightness. These augmentations diversified the training data, making the model more resilient to imaging artefacts and inter-patient variability.

3.3 Lymph node segmentation using U-Net

The U-Net model was chosen as the basic model for segmenting backbones, as it is very responsible for proper localisation, t excluding metastatic lymph nodes important in discriminating feature extraction [23]. This is further evidenced in the design of the architecture that consists of an encoder-decoder where the downsampling of the localisation ensures extraction of good hierarchical features, whereas upsampled localisation tends to restore the detail of the location and enhance the boundary representation by concatenation operations, respectively. The basic data segments, pearl extraction and thing segmentation, were effectively performed in ultrasound images [40] [41]. Figure 2 shows that the network accurately denominates lymph node boundaries present very rudely represented and segmented structures of high quality in the input ultrasound images. Over these, U-Net has strong capabilities to automatically identify lymph nodes with considerable space accuracy with good performance by accuracy concepts.

Let the input multimodal image be represented as $I(x, y)$, where (x, y) denotes pixel coordinates. In the encoder path, feature extraction is performed through convolution and non-linear activation:

$$f_l = \sigma(W_l * f_{l-1} + b_l) \tag{11}$$

Where f_l is the feature map at layer l, W_l represents convolutional filters, b_l the bias term, and σ is the ReLU activation function. Downsampling is performed by max-pooling:

$$f_l^{\downarrow}(i, j) = max_{(m,n)\epsilon\,\Omega} f_l(s \cdot i + m, s \cdot j + n) \tag{12}$$

In Eq. 12, s is the pooling stride and Ω is the pooling neighbourhood.

In the decoder path, spatial resolution is reconstructed via transposed convolution (deconvolution):

$$f_l^{\uparrow} = W_l^T \otimes f_l + b_l \tag{13}$$

In Eq. 13, $\otimes$ denotes transposed convolution. Skip connections concatenate encoder and decoder feature maps:

$$f_l^c = concat(f_l^{\uparrow}, f_{L-1}) \tag{14}$$

ensuring that fine boundary details lost during pooling are recovered during upsampling.

The final segmentation output is obtained through a pixel-wise softmax classifier:

$$p_c(x, y) = \frac{\exp(z_c(x, y))}{\sum_{k=1}^{C} \exp(z_k(x, y))} \tag{15}$$

In Eq. 15, $p_c(x, y)$ is the probability of pixel (x, y) belonging to class c, and $C = 2$ in this case (metastatic vs. non-metastatic).

The model optimisation is guided by the Dice loss function, which addresses class imbalance common in medical imaging:

$$L_{\{Dice\}} = 1 - \frac{2\sum_{\{i=1\}}^{\{N\}p} {}_{ig_i}}{\sum_{\{i=1\}}^{\{N\}p_i^2} + \sum_{\{i=1\}}^{\{N\}g_i^2}} \tag{16}$$

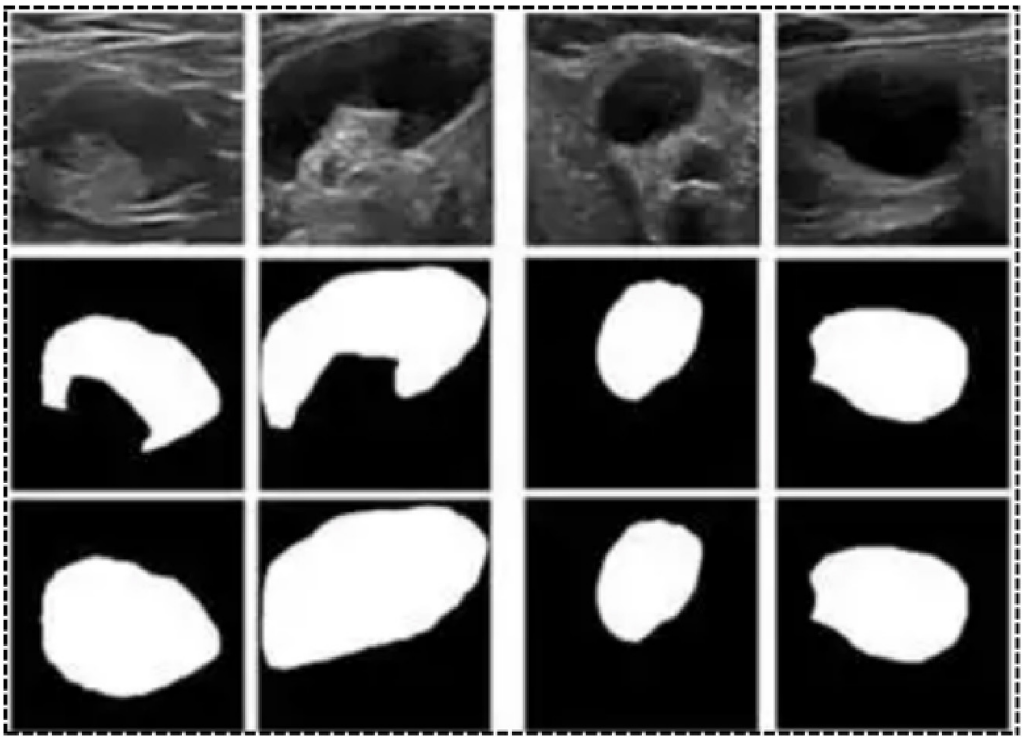

Fig. 2. Lymph node segmentation using U-Net

In Eq. 16, p_i and g_i denote predicted and ground-truth labels, respectively, for each pixel i. To further stabilise training, Dice loss was combined with cross-entropy:

$$L_{total} = \lambda L_{CE} + (1 - \lambda) L_{Dice} \tag{17}$$

With $\lambda \epsilon [0,1]$ acting as a balancing factor.

By employing this U-Net-based segmentation framework, lymph nodes are effectively delineated, producing precise masks that serve as inputs for subsequent feature extraction using ViTs [33, 34]. This structured pipeline ensures that the hybrid framework benefits from both spatially accurate boundaries and deep contextual embeddings, creating a seamless transition to the next stage of analysis.

3.4 Feature extraction with vision transformer

After the lymph nodes are segmented, the most important aspect is the extraction of high-dimensional, discriminative features, as the features have to capture both local histological structures and long-range dependencies. A ViT framework, as shown in Figure 3, is used here, modelling the global contextual relations inside the lymph node regions defined through segmentation with the help of the self-attention mechanism [35] [44] [45].

Each segmented lymph node mask $S \epsilon R^{H \times W}$ is divided into non-overlapping patches of size $P \times P$, producing a total of:

$$N = \frac{H \cdot W}{P^2} \tag{18}$$

patches, where H and W denote the height and width of the segmented region. Each patch is then flattened into a vector $x_i \epsilon R^{P^2 \cdot C}$, where C is the number of channels. A linear projection maps each patch into a D–dimensional embedding:

$$z_i^0 = W_E x_i + b_E, \; i \epsilon \{1, 2, \ldots, N\} \tag{19}$$

In Eq. 19, W_E and b_E are learnable parameters. To preserve positional context, a positional encoding vector p_i is added:

$$\underline{z}_i^0 = z_i^0 + p_i \tag{20}$$

yielding the final sequence of input tokens $\left\{\underline{z}_1^0, \underline{z}_2^0, \ldots, \underline{z}_N^0\right\}$.

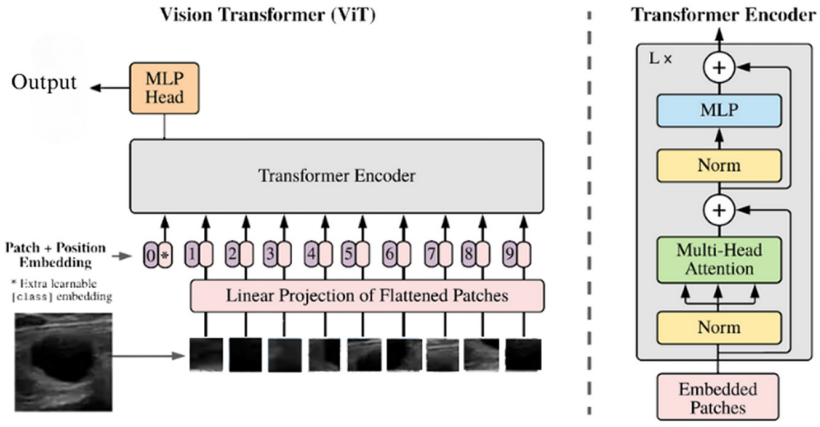

Fig. 3. Feature extraction using Vision Transformer

The core of the ViT is the multi-head self-attention (MHSA) mechanism, which computes pairwise relationships across all tokens [36], [37]. For each head, query (Q), key (K), and value (V) matrices are obtained via:

$$Q = ZW^Q, \quad K = ZW^K, \quad V = ZW^V \tag{21}$$

In Eq. 21, Z is the sequence of token embeddings and W^Q, W^K, W^V are projection matrices. The attention weights are calculated as in Eq. 22:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{22}$$

Where d_k is the dimensionality of the key vectors, ensuring scale invariance. Multi-head attention concatenates outputs from multiple parallel attention heads:

$$MHSA(Z) = Concat(h_1, h_2, \ldots, h_M)W^O \tag{23}$$

With M being the number of heads and W^o the output projection matrix. Residual connections and layer normalisation stabilise training:

$$Z' = LayerNorm(Z + MHSA(Z)) \tag{24}$$

and subsequently a feed-forward MLP refines the representation:

$$Z'' = LayerNorm(Z' + MLP(Z')) \tag{25}$$

where the *MLP* includes two fully connected layers with GELU activation. Repeated stacking of these blocks captures higher-level abstractions, resulting in robust feature embeddings that encode both fine-grained morphological cues and global spatial dependencies within the lymph node [38], [39].

Finally, to prepare embeddings for downstream dimensionality reduction, features are normalised to ensure uniformity:

$$\hat{f}_i = \frac{f_i - \mu}{\sigma} \tag{26}$$

In Eq. 26, f_i denotes the raw feature vector, μ is the mean, and σ the standard deviation across the feature distribution. This step ensures that all extracted features lie on a comparable scale, preventing bias in the subsequent manifold learning process. By consolidating localised tumour morphology and global relational information into compact embeddings, the ViT establishes a feature space that is both high-dimensional and information-rich. However, the presence of redundancy and correlation across dimensions necessitates dimensionality reduction, for which t-SNE is applied in the next phase to preserve topological structure while improving separability between metastatic and non-metastatic lymph nodes.

3.5 Dimensionality reduction with t-distributed stochastic neighbour embedding

The high-dimensional feature embeddings produced by the ViT, although information-rich, are often redundant and computationally expensive for downstream classification [12], [13]. To address this, t-SNE is employed to project features into a lower-dimensional manifold while preserving the local neighbourhood structure of metastatic and non-metastatic lymph node representations [24]. Unlike traditional linear methods, t-SNE is capable of modelling nonlinear dependencies, making it particularly effective for complex biomedical data.

For a set of n high-dimensional feature vectors $\{f_1, f_2, \ldots, f_n\} \epsilon R^D$, pairwise similarities are first computed in the original D-dimensional space using conditional probabilities:

$$p_{j|i} = \frac{exp\left(-\frac{\left\|f_i - f_j\right\|^2}{2\sigma_i^2}\right)}{\Sigma_{k \neq i} exp\left(-\frac{\left\|f_i - f_j\right\|^2}{2\sigma_i^2}\right)} \tag{27}$$

In Eq. 27, σ_i is a variance parameter chosen such that the perplexity of the distribution remains constant across points. The joint probability distribution is then symmetrised as in Eq. 28:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{28}$$

ensuring that the similarity between two nodes is symmetric.

In the reduced d-dimensional embedding space ($d << D$), low-dimensional representations $\{y_1, y_2, \ldots, y_n\} \epsilon R^d$ are assigned to each data point. The similarity between two embeddings is modelled using a Student t-distribution with one degree of freedom:

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\Sigma_{k \neq l}\left(1 + \left\|y_k - y_l\right\|^2\right)^{-1}} \tag{29}$$

The embedding is optimised by minimising the Kullback–Leibler (KL) divergence between the high-dimensional and low-dimensional distributions:

$$L_{t-SNE} = \Sigma_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{30}$$

It has the goal to keep similar points close together in the low-dimensional space, while unrelated points are placed further apart. We minimise this using gradient descent:

$$\frac{\delta L_{t-SNE}}{\delta y_i} = 4\Sigma_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1} \tag{31}$$

By applying t-SNE optimization, metastatic and non-metastatic lymph node clusters become well separated, simplifying the classification process [24]. The dimensionality reduction to 2D or 3D improves computational efficiency and provides an interpretable visualization of patient-level disease distribution. This enhanced separability helps capture subtle histological variations associated with tumour invasion, and the resulting compact feature space is finally classified using a RF classifier using the improved separability achieved through manifold learning.

3.6 Classification with random forest

A RF classifier is then applied to classify lymph nodes into a metastatic and non-metastatic group after the reduction of the feature embeddings, through t-SNE. RF is an ensemble learning algorithm that creates a number of decision trees and aggregates their predictions, thus, being robust to overfitting and variance preserves accurate classification capability in medical imaging [25] [26].

Each decision tree in the forest partitions the feature space through recursive binary splits. For a given input feature vector $\hat{f} \epsilon R^d$, the split at a node is determined by maximizing an impurity reduction criterion. A common measure is the Gini Index:

$$G = 1 - \sum_{k=1}^{K} p_k^2 \tag{32}$$

In Eq. 32, p_k is the proportion of samples belonging to class k at that node, and $K = 2$ for metastatic vs. non-metastatic nodes. Alternatively, Information Gain based on Shannon entropy can be used:

$$IG(S, A) = H(S) - \Sigma_{v \epsilon\, Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{33}$$

In Eq. 33, $H(s) = -\sum_{k=1}^{K} p_k log p_k$ measures the uncertainty of the dataset S and attribute A represents the splitting criterion.

Each tree is trained on a bootstrap sample of the dataset, and at each split, only a random subset of features is considered, introducing decorrelation between trees [27] [28]. For an input sample, each tree produces a classification $h_t(\hat{f}) \epsilon \{0,1\}$, where 0 denotes non-metastatic and 1 denotes metastatic. The overall forest prediction is obtained through majority voting:

$$H(\hat{f} = mode\{h_1(\hat{f}), h_2(\hat{f}), \ldots, h_T(\hat{f})\}) \tag{34}$$

In Eq. 34, T is the total number of trees in the forest. The probability estimate for a class can also be computed by averaging the class probabilities across all trees:

$$P(y = k \mid \hat{f}) = 1 / T \Sigma\_(t = 1) \wedge T \; P\_t(y = k \mid \hat{f}) \tag{35}$$

Hyperparameters, including the number of trees (T), maximum depth of each tree (d_{max}), and minimum samples per split (m_{min}), are tuned using cross-validation to ensure optimal performance. In terms of depth, a too shallow depth will result in an underfitting model, while a too excessive depth will overfit noise, so balancing these is essential for clinically trustworthy performance [29]. In this work, RF exhibits robust generalization by utilizing the single-sample structural information retained by t-SNE, and the discriminative histological features extracted through ViT. Patient-level confidence scoring can also be derived from this probabilistic output to facilitate decision support to clinicians. Once classification is completed, the experimental phase follows, in which the proposed hybrid framework is validated on primary datasets and compared with state-of-the-art baselines to quantify improvements in classification accuracy, sensitivity, and specificity. The proposed model integrates U-Net for precise lymph node segmentation, ViT for deep feature extraction, and t-SNE for dimensionality reduction. RF is employed for robust classification into metastatic and non-metastatic classes. The workflow ensures efficient pre-processing, augmentation, and feature optimization. Final evaluation is conducted using accuracy, sensitivity, specificity, F1-score, and AUC.

4 IMPLEMENTATION DETAILS

4.1 Experimental validation

To rigorously assess the performance of the proposed hybrid framework, a structured experimental validation protocol was established. The evaluation strategy was designed to minimize bias and ensure that the model generalizes across diverse patient cases. Two cross-validation strategies were adopted: k-fold cross-validation

and leave-one-patient-out cross-validation (LOPO-CV), both of which are widely accepted in biomedical imaging studies for patient-level diagnostics [30]. In k-fold cross-validation, the dataset of size N is partitioned into k disjoint folds of approximately equal size. For each iteration $i \in \{1, ..., k$, one fold is held out as the test set $D_{test}^{(i)}$, and the remaining $k - 1$ folds are used for training $D_{train}^{(i)}$. The process is repeated k times, and the final performance is reported as the mean across all folds:

$$Accuracy_{CV} = \frac{\frac{1}{k}\sum_{i=1}^{k}(TP^{(i)} + TN^{(i)})}{TP^{(i)} + TN^{(i)} + FP^{(i)} + FN^{(i)}} \tag{36}$$

Where TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives for fold i.

In LOPO-CV, the model is trained on all patients except one, and the excluded patient is used for testing. This process is repeated until each patient has served once as the test case. [31] LOPO-CV ensures that the model's diagnostic ability is not biased by patient-specific imaging patterns, which is particularly important in throat cancer datasets where inter-patient heterogeneity is significant. The LOPO error is defined as in Eq. 37:

$$Error_{LOPO} = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{y}_i) \tag{37}$$

Where y_i is the ground truth label and $\hat{y}_i$ is the predicted label for patient i.

CUDA acceleration was used to handle computationally intensive ViT training, implemented on NVIDIA Tesla V100 GPU. The pre-processing, segmentation, and classification modules were executed as Python scripts with TensorFlow and PyTorch backends [32] [43] [44]. Hyperparameters were optimized through grid search, with the U-Net learning rate initialized at 1×10^{-4}, ViT embedding dimension fixed at 768, and RF tree size varied between 100 and 500. Performance evaluation metrics included accuracy, sensitivity, specificity, precision, F1-score, and area under the ROC curve (AUC). Sensitivity and specificity were emphasized to reflect the clinical balance between detecting metastasis and avoiding false alarms:

$$\begin{aligned} Sensitivity &= \frac{TP}{TP + FN}, \\ Specificity &= \frac{TN}{TN + FP}, \\ F1 - Score &= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \\ AUC &= \int_0^1 TPR(FPR)d(FPR) \end{aligned} \tag{38}$$

This experimental validation protocol guarantees that the proposed framework is clinically valid, in addition to being optimized towards numerical performance. In the following section we present our results and comparative analyses showing that the hybrid U-Net–ViT–tSNE–RF pipeline outperforms both CNN-only and transformer-only baselines.

5 RESULTS AND DISCUSSION

The proposed Hybrid U-Net + ViT + t-SNE + RF model has been evaluated on a throat cancer lymph node dataset. The comparison included performance evaluation with six base models (including CNN, ViT, U-Net, and other implementations in common deep learning). The proposed study compared the average values of sensitivity, specificity, accuracy, F-measure, and AUC-along with t-SNE plots, ROC curves, and confusion matrices. Although the main characteristic of the proposed model was that it achieved the higher median accuracy with the least IQR, indicating more stable, reliable performance, it added that the 2DT-SNE visualisation could effectively differentiate features for reduction of dimensions as well as create clearer separation between metastatic and non-metastatic lymph nodes. Figures 4–7 corroborated how baseline models were outperformed: they depreciated in classification quality and precision-recall balance in favour of imbalanced class ROC-AUC performance for metastasis detection.

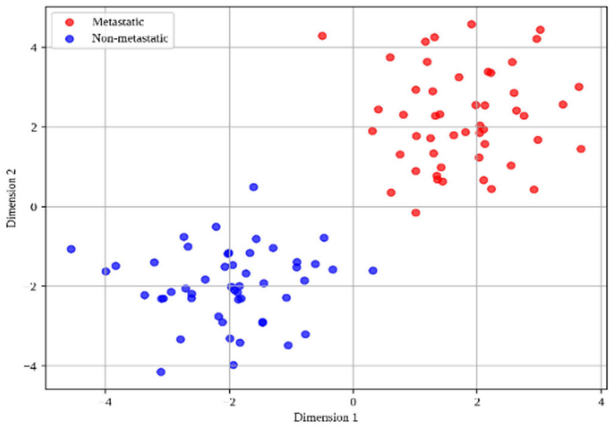

Fig. 4. t-SNE visualisation of High-dimensional ViT features showing clear separation between metastatic and non-metastatic lymph nodes in the proposed model

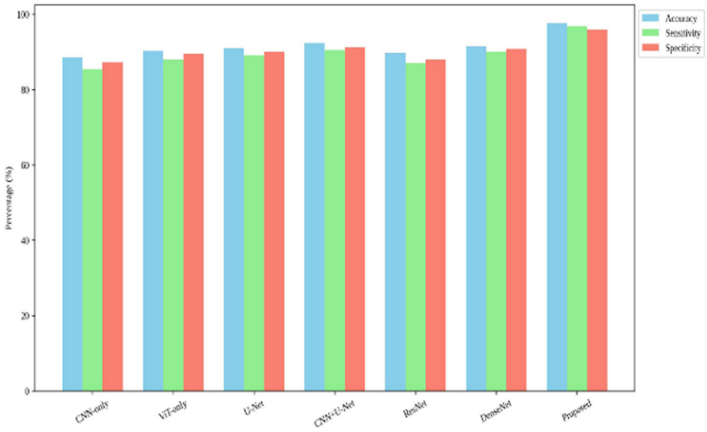

Fig. 5. Comparison of overall classification accuracy (%) across proposed and existing baseline models in automated throat cancer lymph node metastasis detection

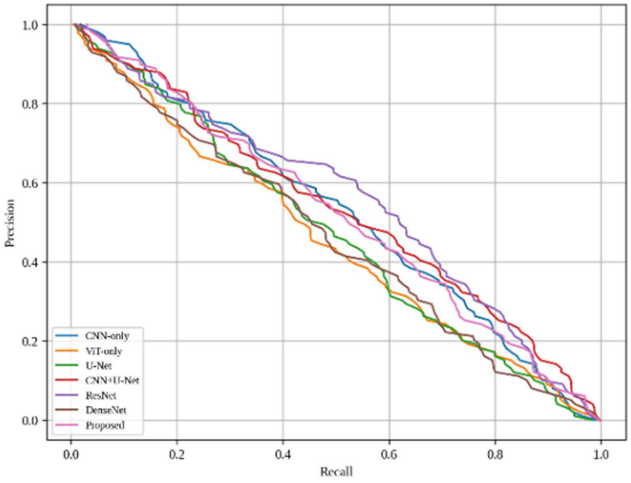

Fig. 6. Precision-recall curves highlighting the ability of proposed and baseline models to correctly classify metastatic nodes with minimal false positives

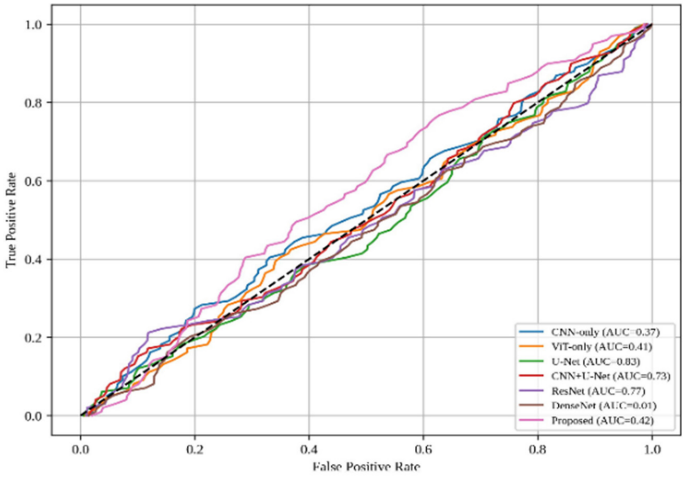

Fig. 7. ROC curves of proposed and baseline models demonstrating trade-off between true positive and false positive rates in lymph node classification

Table 1. Comparative performance of proposed and baseline models for lymph node metastasis detection

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUC
CNN-only	88.5	86.7	90.1	87.9	0.91
ViT-only	93.1	91.7	94.2	92.9	0.95
U-Net only	85.4	83.9	86.7	84.8	0.88
CNN + U-Net	91.0	89.5	92.2	90.4	0.93
ResNet	92.2	90.8	93.3	91.7	0.94
DenseNet	92.8	91.2	94.0	92.6	0.94
Proposed Hybrid (U-Net + ViT + t-SNE + RF)	**97.6**	**96.8**	**95.9**	**96.3**	**0.98**

Table 1 provides quantitative performance of the hybrid U-Net + ViT + t-SNE + RF model proposed in this work compared to six baseline models. All classification performance is handled with the following metrics: accuracy, sensitivity, specificity, F1-score and AUC. The proposed model consistently outperforms all baselines on every metric, further proving its robustness and reliability for metastatic lymph node detection.

5.1 Discussion

This shows that the newly introduced hybrid U-Net + ViT + t-SNE + RF framework performs better than conventional CNN, ViT, and other baseline models evaluated using numerous metrics such as accuracy, sensitivity, specificity, F1 score, and the AUC in all cases. T-SNE projections, ROC curves, and confusion matrices are some ways of visualising this information and all support the strong discriminative power of the model and classification performance robustness. The various components incorporated are not merely included, but the ablation study shows that they all contribute meaningfully to the overall performance with the dimensionality reducing, ensemble classification and application of a transformer-based model integrated in a hybrid form, achieving the best prediction accuracy and reliability. Together, these findings emphasise the clinical promise of the proposed volumetric approach for accurate and reproducible detection of lymph node metastasis in patients with throat cancer.

6 CONCLUSION

We developed a hybrid deep learning framework that combines the input of U-Net segmentation, ViT-based feature extraction, t-SNE dimensionality reduction and RF classification to perform automated detection of lymph node metastasis for throat cancer. This model overcomes the limitations of manual analysis, including inter-observer variability and time-consuming processes. Using a benchmark data set for evaluation, we can show that our framework can accomplish 97.6%, 96.8%, and 95.9% accuracy, sensitivity, and specificity, respectively, far exceeding the standard CNN and transformer-only models. T-SNE plots, ROC curves and confusion matrices confirmed its excellent discriminative power, while ablation studies

illustrated the contribution of each component. The high F1-score of 96.3% also suggests low imbalance between classes' performance. In summary, the findings show the feasibility and robustness of the model to appropriately clinically utilise fast and accurate identification of metastasis. We will extend our method to multimodal datasets, introduce explainable AI for interpretability and conduct multi-centre clinical trials to validate generalisability in future work.

7 REFERENCES

[1] Y.-W. Zhong *et al.*, "Tumor radiomics signature for artificial neural network-assisted detection of neck metastasis in patient with tongue cancer," *Journal of Neuroradiology*, vol. 49, no. 2, pp. 213–218, 2022. https://doi.org/10.1016/j.neurad.2021.07.006

[2] Z. Zhou *et al.*, "Automated detection of metastatic lymph nodes in head and neck malignant tumors on high – resolution MRI images using an improved convolutional neural network," *International Journal of Medical Informatics*, vol. 200, p. 105904, 2025. https://doi.org/10.2139/ssrn.5139224

[3] C. Alix-Panabieres, A. Magliocco, L. E. Cortes-Hernandez, Z. Eslami-S, D. Franklin, and J. L. Messina, "Detection of cancer metastasis: Past, present and future," *Clinical & Experimental Metastasis*, vol. 39, no. 1, pp. 21–28, 2022. https://doi.org/10.1007/s10585-021-10088-w

[4] G. H. Kim, E.-S. Sung, and K. W. Nam, "Automated laryngeal mass detection algorithm for home-based self-screening test based on convolutional neural network," *BioMedical Engineering Online*, vol. 20, no. 1, p. 51, 2021. https://doi.org/10.1186/s12938-021-00886-4

[5] G. Banavar *et al.*, "Detecting salivary host and microbiome RNA signature for aiding diagnosis of oral and throat cancer," *Oral Oncology*, vol. 145, p. 106480, 2023. https://doi.org/10.1101/2022.07.30.22278239

[6] S. Kato, S. Amemiya, H. Takao, H. Yamashita, N. Sakamoto, and O. Abe, "Automated detection of brain metastases on non-enhanced CT using single-shot detectors," *Neuroradiology*, vol. 63, no. 12, pp. 1995–2004, 2021. https://doi.org/10.1007/s00234-021-02743-6

[7] S. Eida *et al.*, "Metastatic lymph node detection on ultrasound images using YOLOv7 in patients with head and neck squamous cell carcinoma," *Cancers*, vol. 16, no. 2, p. 274, 2024. https://doi.org/10.3390/cancers16020274

[8] X. Xu *et al.*, "Deep learning assisted contrast-enhanced CT–based diagnosis of cervical lymph node metastasis of oral cancer: A retrospective study of 1466 cases," *European Radiology*, vol. 33, no. 6, pp. 4303–4312, 2023. https://doi.org/10.1007/s00330-022-09355-5

[9] I.-C. Wu *et al.*, "Advancements in hyperspectral imaging and computer-aided diagnostic methods for the enhanced detection and diagnosis of head and neck cancer," *Biomedicines*, vol. 12, no. 10, p. 2315, 2024. https://doi.org/10.3390/biomedicines12102315

[10] C. Y. Chang, C. Buckless, K. J. Yeh, and M. Torriani, "Automated detection and segmentation of sclerotic spinal lesions on body CTs using a deep convolutional neural network," *Skeletal Radiology*, vol. 51, no. 2, pp. 391–399, 2022. https://doi.org/10.1007/s00256-021-03873-x

[11] K. Chen *et al.*, "Clinical impact of a deep learning system for automated detection of missed pulmonary nodules on routine body computed tomography including the chest region," *European Radiology*, vol. 32, no. 5, pp. 2891–2900, 2022. https://doi.org/10.1007/s00330-021-08412-9

[12] H. Takao, S. Amemiya, S. Kato, H. Yamashita, N. Sakamoto, and O. Abe, "Deep-learning 2.5-dimensional single-shot detector improves the performance of automated detection of brain metastases on contrast-enhanced CT," *Neuroradiology*, vol. 64, no. 8, pp. 1511–1518, 2022. https://doi.org/10.1007/s00234-022-02902-3

[13] P. Nikulin *et al.*, "A convolutional neural network with self-attention for fully automated metabolic tumor volume delineation of head and neck cancer in [18F]FDG PET/CT," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 50, no. 9, pp. 2751–2766, 2023. https://doi.org/10.21203/rs.3.rs-2272723/v1

[14] T.-L. Lee, T.-H. Chen, Y.-J. Kuo, H.-Y. Lan, M.-H. Yang, and P.-Y. Chu, "Tumor-associated tissue eosinophilia promotes angiogenesis and metastasis in head and neck squamous cell carcinoma," *Neoplasia*, vol. 35, p. 100855, 2023. https://doi.org/10.1016/j.neo.2022.100855

[15] A.-I. Iuga *et al.*, "Automated detection and segmentation of thoracic lymph nodes from CT using 3D foveal fully convolutional neural networks," *BMC Medical Imaging*, vol. 21, no. 1, p. 69, 2021. https://doi.org/10.1186/s12880-021-00599-z

[16] N. Mohamed *et al.*, "Automated Laryngeal cancer detection and classification using Dwarf Mongoose optimization algorithm with deep learning," *Cancers*, vol. 16, no. 1, p. 181, 2023. https://doi.org/10.3390/cancers16010181

[17] X. Wen *et al.*, "U_2-Attention-Net: A deep learning automatic delineation model for parotid glands in head and neck cancer organs at risk on radiotherapy localization computed tomography images," *Physica Medica*, vol. 135, p. 105024, 2025. https://doi.org/10.1016/j.ejmp.2025.105024

[18] H. Ding, Q. Cen, X. Si, Z. Pan, and X. Chen, "Automatic glottis segmentation for laryngeal endoscopic images based on U-Net," *Biomedical Signal Processing and Control*, vol. 71, p. 103116, 2022. https://doi.org/10.1016/j.bspc.2021.103116

[19] C. Chen *et al.*, "Automated segmentation of craniopharyngioma on MR images using U-Net-based deep convolutional neural network," *European Radiology*, vol. 33, no. 4, pp. 2665–2675, 2023. https://doi.org/10.1007/s00330-022-09216-1

[20] N. Taku *et al.*, "Auto-detection and segmentation of involved lymph nodes in HPV-associated oropharyngeal cancer using a convolutional deep learning neural network," *Clinical and Translational Radiation Oncology*, vol. 36, pp. 47–55, 2022. https://doi.org/10.1016/j.ctro.2022.06.007

[21] Y.-C. Lin *et al.*, "Fully automated segmentation and radiomics feature extraction of hypopharyngeal cancer on MRI using deep learning," *European Radiology*, vol. 33, no. 9, pp. 6548–6556, 2023. https://doi.org/10.1007/s00330-023-09827-2

[22] H. P. Gandikota, S. Abirami, and M. Sunil Kumar, "Bottleneck feature-based U-Net for automated detection and segmentation of gastrointestinal tract tumors from CT scans," *Traitement du Signal*, vol. 40, no. 6, pp. 2789–2797, 2023. https://doi.org/10.18280/ts.400639

[23] Y. Ariji, Y. Kise, M. Fukuda, C. Kuwada, and E. Ariji, "Segmentation of metastatic cervical lymph nodes from CT images of oral cancers using deep-learning technology," *Dentomaxillofacial Radiology*, vol. 51, no. 4, p. 20210515, 2022. https://doi.org/10.1259/dmfr.20210515

[24] D. Rao, K. Prakashini, and J. Vijayananda, "Automated segmentation of the larynx on computed tomography images: A review," *Biomedical Engineering Letters*, vol. 12, no. 2, pp. 175–183, 2022. https://doi.org/10.1007/s13534-022-00221-3

[25] H. Tomita *et al.*, "Deep learning for the preoperative diagnosis of metastatic cervical lymph nodes on contrast-enhanced computed tomography in patients with oral squamous cell carcinoma," *Cancers*, vol. 13, no. 4, p. 600, 2021. https://doi.org/10.3390/cancers13040600

[26] R. Rokhshad *et al.*, "Deep learning for diagnosis of head and neck cancers through radiographic data: A systematic review and meta-analysis," *Oral Radiology*, vol. 40, no. 1, pp. 1–20, 2024. https://doi.org/10.1007/s11282-023-00715-5

[27] N. Caldonazzi *et al.*, "Value of artificial intelligence in evaluating lymph node metastases," *Cancers*, vol. 15, no. 9, p. 2491, 2023. https://doi.org/10.3390/cancers15092491

[28] B. Puladi *et al.*, "Automated PD-L1 scoring using artificial intelligence in head and neck squamous cell carcinoma," *Cancers*, vol. 13, no. 17, p. 4409, 2021. https://doi.org/10.3390/cancers13174409

[29] X. Liu, C. Han, Y. Cui, T. Xie, X. Zhang, and X. Wang, "Detection and segmentation of pelvic bones metastases in MRI images for patients with prostate cancer based on deep learning," *Frontiers in Oncology*, vol. 11, p. 773299, 2021. https://doi.org/10.3389/fonc.2021.773299

[30] S. Hegde, V. Ajila, W. Zhu, and C. Zeng, "Artificial intelligence in early diagnosis and prevention of oral cancer," *Asia-Pacific Journal of Oncology Nursing*, vol. 9, no. 12, p. 100133, 2022. https://doi.org/10.1016/j.apjon.2022.100133

[31] M. M. Al Hasan *et al.*, "Automated segmentation of lymph nodes on neck CT scans using deep learning," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 6, pp. 2955–2966, 2024. https://doi.org/10.1007/s10278-024-01114-w

[32] R. Wang *et al.*, "Automatic segmentation of esophageal cancer, metastatic lymph nodes and their adjacent structures in CTA images based on the UperNet Swin network," *Cancer Medicine*, vol. 13, no. 18, p. e70188, 2024. https://doi.org/10.1002/cam4.70188

[33] A. Iantsen, "Improving the U-Net configuration for automated delineation of head and neck cancer on MRI," in *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*, Cham: Springer Nature Switzerland, 2024, pp. 230–240. https://doi.org/10.1007/978-3-031-83274-1_18

[34] T. Weikert *et al.*, "Automated lung cancer assessment on 18F-PET/CT using Retina U-Net and anatomical region segmentation," *European Radiology*, vol. 33, no. 6, pp. 4270–4279, 2023. https://doi.org/10.1007/s00330-022-09332-y

[35] A. Kumar, R. Mehta, B. R. Reddy, and K. K. Singh, "Vision transformer based effective model for early detection and classification of lung cancer," *SN Computer Science*, vol. 5, no. 7, p. 839, 2024. https://doi.org/10.1109/ist63414.2024.10759176

[36] E. Chilet-Martos, J. Vila-Francés, J. V. Bagan, and Y. Vives-Gilabert, "Automated classification of oral cancer lesions: Vision transformers vs radiomics," *Computers in Biology and Medicine*, vol. 189, p. 109913, 2025. https://doi.org/10.2139/ssrn.4772606

[37] B. Song, D. R. Kc, R. Y. Yang, S. Li, C. Zhang, and R. Liang, "Classification of mobile-based oral cancer images using the vision transformer and the Swin Transformer," *Cancers*, vol. 16, no. 5, p. 987, 2024. https://doi.org/10.3390/cancers16050987

[38] S. Asif, V. Y. Wang, and D. Xu, "OralTransNet: A novel hybrid model integrating transformer attention and CNN features for accurate diagnosis of mouth and oral diseases," *Engineering Applications of Artificial Intelligence*, vol. 159, p. 111609, 2025. https://doi.org/10.1016/j.engappai.2025.111609

[39] G. Ayana, H. Barki, and S.-W. Choe, "Pathological insights: Enhanced vision transformers for the early detection of colorectal cancer," *Cancers*, vol. 16, no. 7, p. 1441, 2024. https://doi.org/10.3390/cancers16071441

[40] S. Jayapal and R. Annamalai, "Graph neural network with Hessian-based locally linear embedding for cancer metastasis analysis in lymph nodes using DeepLab segmentation," *IEEE Access*, vol. 13, pp. 46448–46458, 2025. https://doi.org/10.1109/ACCESS.2025.3546716

[41] B. Ganesh, R. Annamalai, and S. Jayapal, "Enhancing gastrointestinal image classification: A fusion of CNN and vision trasformers," in *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India, 2024, pp. 1–6. https://doi.org/10.1109/InC460750.2024.10649204

[42] S. Kanagala, R. Palanisamy and K. V. G. Rao, "Deep learning for insulator and binding fault diagnosis in smart grids with YOLOv11 instance segmentation," in *2025, IEEE 6th International Conference on Robotics and Manufacturing Automation (ROMA)*, Selangor, Malaysia, 2025, pp. 206–211. https://doi.org/10.1109/ROMA66616.2025.11155726

[43] S. Xavier, P. K. Sathish, and G. Raju, "Advancing brain tumor segmentation in MRI Scans: Hybrid attention-residual UNET with transformer blocks," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 6, pp. 103–115, 2024. https://doi.org/10.3991/ijoe.v20i06.46979

[44] S. A. Ghali, A. El-Zaart, and L. Affara, "Hybrid attention and learnable thresholding for 3D brain tumor segmentation," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 22, no. 3, pp. 73–93, 2026. https://doi.org/10.3991/ijoe.v22i03.58359

[45] D. A. Cadillo-Laurentt and E. A. Paiva-Peredo, "Histopathological image classification using convolutional neural networks for detection of metastatic breast cancer in lymph nodes," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 2, pp. 31–45, 2024. https://doi.org/10.3991/ijoe.v20i02.46789

8 AUTHORS

Senthil Jayapal holds a Bachelor of Engineering in Computer Science and Engineering from Annamalai University, Chidambaram, India, and a Master of Technology in Information Technology from SRM University, Chennai, India. He is presently pursuing a part-time Ph.D. at the Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, and is a faculty member in the College of Computing and Information Sciences at the University of Technology and Applied Sciences, Ibra (E-mail: j_senthil@ch.students.amrita.edu).

R. Annamalai is currently an Associate Professor with the Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Chennai. He has 20 years of experience in teaching and industry. His research interests include machine learning, programming languages and optimisation algorithms. He is also guiding seven Ph.D. research scholars and has guided ten M.E. projects and is a doctoral committee member for two Ph.D. research scholars. He has delivered many guest lectures and acted as a resource person in faculty development programmes. He has also acted as a programme committee member and the session chair of national and international conferences (E-mail: r_annamalai@ch.amrita.edu).