IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Enhancing Thyroid Nodule Assessment with UTV-ST Swin Kansformer: A Multimodal Approach to Predict Invasiveness

YuFang Zhao[1], Yue Li[2], Yanjing Zhang[1], Xiaohui Yan[1] , GuoLin Yin[1] ,and Liping Liu[1]

[1]Department of Ultrasound Intervention, First Hospital of Shanxi Medical University, Taiyuan,Shanxi, China

[2] School of Electronic and Information Engineering, TianGong University,Tianjin,China

Corresponding author: Liping Liu (e-mail: liuliping1600@sina.com).

This study was supported by Key Research and Development Program of Science and Technology, Department of Shanxi Province (201903D321190) and Ministry of Human and Social Affairs Program for the selection of Science and Technology activities for Returned Overseas Scholar (2016-366)

ABSTRACT Assessing the invasiveness of thyroid nodules, particularly whether they have metastasized to lymph nodes, is crucial for guiding treatment decisions. Current diagnostic methods, including ultrasound imaging, are limited by operator dependence and interpretative variability, complicating accurate evaluation of nodule invasiveness. To address these limitations, this study introduces the UTV-ST Swin Transformer, a deep learning model that combines ultrasound video data with standardized clinical information to predict the invasiveness of thyroid nodules. The model classifies nodules into three categories: non-invasive, central lymph node metastasis (CLNM), and central plus lateral lymph node metastasis (CLNM+LLNM). By analyzing ultrasound video features using the Video Swin Transformer and clinical data using a text analysis module based on the KAN network, and then fusing these features, the model achieves a classification accuracy of 82.1% and an average AUC of 94.2%. These results surpass the performance of traditional methods, particularly in distinguishing different degrees of invasiveness, even under noisy conditions. This study highlights the potential of the UTV-ST Swin Transformer model in improving the accuracy of thyroid nodule assessment, reducing reliance on operator expertise, and providing a more consistent and automated method for evaluating nodule invasiveness.

INDEX TERMS Papillary thyroid carcinoma (PTC), Cervical lymph node metastasis, Ultrasound imaging, Kolmogorov-Amold Network, Video Abnormal Detection

I. INTRODUCTION

Thyroid cancer is the most common endocrine malignancy, with its incidence continually increasing worldwide, particularly the significant rise in papillary thyroid carcinoma (PTC) [1]. Patients with PTC and follicular thyroid carcinoma (FTC) have a 5-year overall survival rate exceeding 98%; however, once it progresses to anaplastic thyroid carcinoma (ATC), the median overall survival drops to only a few weeks to months [2]. Late detection and diagnosis of PTC increase the likelihood of dedifferentiation, potentially progressing to ATC. Cervical lymph node metastasis is the primary pathway of thyroid cancer spread and a major factor contributing to local recurrence [3,4]. Patients with lymph node metastasis have an average recurrence rate of 22%, whereas those without metastasis have a recurrence rate of only 2% [5,6]. Once distant metastasis occurs, the 10-year overall survival rate decreases to below 50% [6]. While prophylactic lateral neck lymph node dissection can reduce the risk of missing metastatic lymph nodes, it involves extensive surgical scope, increases the incidence of complications, and reduces patients' quality of life. Conversely, not performing dissection may miss some metastatic cervical lymph nodes, affecting patients' long-term survival [7,8]. Therefore, accurate preoperative assessment of thyroid nodule invasiveness is crucial for determining the surgical extent of lateral neck lymph node

dissection, formulating individualized and rational treatment plans, and improving patient prognosis [9,10].

Traditional ultrasound image classification methods are mainly divided into contour-based and feature-learning-based approaches. Although these methods perform reasonably well in classification tasks, manual feature extraction processes are complex, and results are sometimes suboptimal. Some researchers have used traditional machine learning algorithms [11-13] to analyze ultrasound images, but this requires manual design of feature extraction algorithms, making it difficult to apply to large-scale medical data. In contrast, deep learning constructs deep convolutional neural networks for big data training, featuring automatic feature learning and strong robustness. Building upon an in-depth study of existing mainstream deep learning methods and our team's previous research on thyroid nodule classification [14], we propose a novel and reliable classification model — the UTV-ST Swin Transformer. This method employs a multimodal learning framework that combines video data and standardized clinical information, focusing on the nodule regions in videos to provide recommendations for nodule classification and treatment decisions.

II. Related Work

A. Ultrasound thyroid classification method

Despite the diversity of thyroid nodules, research typically focuses on the binary classification of benign and malignant cases. Early studies primarily relied on machine learning and statistical methods to analyze imaging, clinical, and ultrasound features individually, with few studies addressing the integration of multimodal data. To address this limitation, a nomogram model was developed that combines ultrasound and clinical data to predict the risk of central lymph node metastasis in PTC patients. This model utilizes grayscale imaging and six relevant features, but its prediction accuracy is compromised due to the absence of multimodal ultrasound data. He et al. [16] retrospectively collected B-type ultrasound data from patients at two centers, developing 28 models using seven machine learning algorithms combined with four types of imaging data: B-US, B-US + CDFI + RTE, CEUS, and B-US + CDFI + RTE + CEUS. The results showed that the diagnostic performance of the machine learning model was comparable to senior radiologists but outperformed junior radiologists. Hai Du et al. [17] conducted a retrospective study involving 1,076 thyroid nodules from 817 patients across three institutions. They extracted radiomics and deep learning features from ultrasound images, constructing radiomics features (Rad_sig) and deep learning features (DL_sig). Feature selection was carried out using Pearson correlation analysis and LASSO regression. Additionally, clinical ultrasound semantic features (C_US_sig) were derived from clinical data and ultrasound semantic information. The final model was a nomogram combining these three feature types.

With the advancement of deep learning technology, particularly convolutional neural networks (CNNs), significant progress has been made in using computer vision tasks to diagnose thyroid nodules. Miribi Rho et al. [18] evaluated the performance of deep CNNs in distinguishing benign and malignant thyroid nodules smaller than 10 mm, comparing CNN diagnostic performance with that of radiologists. The results demonstrated that CNNs trained on thyroid nodules larger than 10 mm performed better than radiologists in diagnosing and classifying smaller nodules, especially those $\leq$ 5 mm. Chen Chen et al. [19] conducted a multicenter retrospective study using ultrasound images from four hospitals. They developed a CNN model to classify thyroid nodules into solid vs. non-solid and benign vs. malignant categories, with the Inception-ResNet architecture achieving the highest AUC (0.94). Na Zhang et al. [20] proposed a hybrid deep learning model combining ultrasound and infrared thermal imaging to classify thyroid nodules, offering a promising non-invasive diagnostic method for assessing malignancy. Liu et al. [21] developed a novel deep learning model, DualSwinThyroid, which integrates multimodal ultrasound imaging data and clinical information to predict cervical lymph node metastasis in PTC patients. This model provides early and accurate identification, enabling better strategic decisions regarding surgical interventions for high-risk PTC patients.

B. Swin Transformer

Swin Transformer is a deep learning model derived from the Transformer architecture [22], demonstrating great potential in extracting features from various data types [23-26]. It constructs a hierarchical structure by dividing the input into non-overlapping windows and applying self-attention mechanisms within each window. To capture broader contextual information, Swin Transformer incorporates a shifted window process in subsequent layers. Studies have shown that Swin Transformer outperforms traditional CNN architectures in many applications [25,26].

In video understanding tasks, the Video Swin Transformer extends the two-dimensional Swin Transformer into three dimensions, allowing it to directly process video data [26]. Compared to commonly used temporal models like RNNs, Swin Transformer-based models effectively address the vanishing gradient problem. Moreover, by treating each 3D block of a video as a token, they provide a more compact and efficient representation for video processing. In contrast, RNNs require a large number of input tokens to achieve similar video representations [26].

C. Kolmogorov-Amold Network

Kolmogorov-Amold Networks (KANs) are based on the Kolmogorov-Arnold representation theorem, which states that any multivariate continuous function can be represented as a sum of compositions of univariate continuous functions [37]. This provides both the existence and a constructive method for such representations.Unlike MLPs, which apply activation functions to the nodes, KANs place learnable activation functions on the edges between nodes. These activation functions are based on univariate B-splines, which replace the fixed activation functions (e.g., ReLU or Sigmoid) used in MLPs, allowing KANs to better model highly nonlinear relationships.Similar to MLPs, a k-layer KAN can be described as a nested structure of multiple KAN layers, as shown in Equation (1):

$$KAN(Z) = (\Phi_{L-1} \circ \Phi_{L-2} \circ \cdots \circ \Phi_1 \circ \Phi_0)Z. \quad (1)$$

KANs consist of interconnected edges with learnable B-spline activation functions, which evolve during training by dynamically adjusting the number of grid points. This flexibility improves the model's accuracy and its ability to capture complex data patterns. By eliminating linear layers and using parameterized activation functions at the edges, KANs offer a more expressive and interpretable framework for modeling complex data relationships.

D. Model Fusion

In multimodal learning, model fusion strategies mainly include early fusion, slow fusion, and late fusion, each differing in data integration methods [27]. Karpathy et al. [28] compared these fusion strategies, particularly in the context of capturing temporal and spatial dependencies in video understanding. Their research indicates that slow fusion, which acquires global information at a higher level, outperforms the alternative schemes of early and late fusion. Feichtenhofer et al. [29] used pre-trained neural networks to extract features from different data sources and performed late fusion using CNNs. Shoukat et al. [30] employed methods such as linear regression to perform weighted averaging of scores from different models to achieve late fusion. When dealing with multimodal sub-models of varying complexity, the VLMo model [31] proposed a staged training method: first training the more complex models, freezing their weights, and then training the simpler models to capture complementary information.

Despite the progress achieved by existing methods, shortcomings remain in practical applications. Building on our team's previous research, this paper proposes a novel classification method for thyroid nodule invasiveness, further deepening the multimodal fusion of ultrasound image features and standardized text features. We extended the application of ultrasound images to short video analysis and optimized the model's ability to classify invasive nodules by incorporating validated feature fusion strategies. Experimental results demonstrate that the proposed method surpasses existing approaches in both classification accuracy and model robustness, effectively enhancing the prediction of thyroid nodule invasiveness.

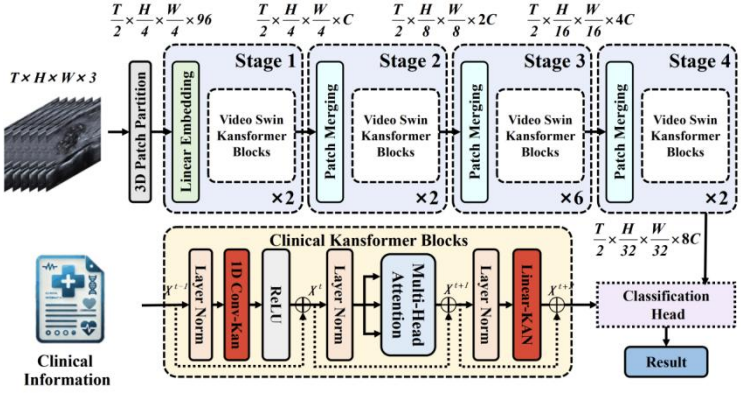
III. UThyroV-ST Swin Kansformer

A. Overall Architecture

UTV-ST Swin Kansformer is a model specifically designed to predict thyroid nodule metastasis based on ultrasound videos. The architecture of the model is shown in Figure 1. The "UTV" in the name stands for "Ultrasound Thyroid Video," indicating that ultrasound videos are the primary input data. "ST" refers to "Standardized Text," highlighting the inclusion of standardized patient text data to enhance the model's interpretability and accuracy. The "Swin Kansformer" is an innovative enhancement that combines the Video Swin Transformer with the KAN network, creating a robust multimodal learning framework.

A successful multimodal predictive model must effectively extract complementary features from each data source and integrate information across multiple modalities. The UTV-ST Swin Kansformer consists of two main components: the Video Swin Kansformer, which processes video data, and the Clinical Kansformer, which processes standardized text data. To combine the outputs from these components, the model uses a staged slow fusion method in the Classification Head, a technique proven effective in previous studies [32].

FIGURE 1. Overall architecture of UThyroV-ST Swin Kansformer



B. .Video Swin Kansformer

The Video Swin Kansformer Blocks are formed by combining the Video Swin Transformer [33] and the KAN network, specifically designed for thyroid ultrasound detection in video data, as shown in Figure 2. These blocks utilize multi-head self-attention (MSA) modules based on 3D shifted windows from the Video Swin Transformer and replace the fully connected MLP layers with the KAN network. Specifically, each Video Swin Kansformer Block includes an MSA module based on 3D

shifted windows, followed by a KAN network. Layer Normalization is applied before each module, and residual connections are incorporated after each module. The computational formula for the Video Swin Kansformer Blocks is provided in Equation (2).

$$\begin{aligned} \hat{Z}^{l} &= 3DW\text{-}MSA(LN(Z^{l-1}))\oplus Z^{l-1} \\ Z^{l} &= KAN(LN(\hat{Z}^{l}))\oplus \hat{Z}^{l} \\ \hat{Z}^{l+1} &= 3DSW\text{-}MSA(LN(Z^{l}))\oplus Z^{l} \\ Z^{l+1} &= KAN(LN(\hat{Z}^{l+1}))\oplus \hat{Z}^{l+1} \end{aligned} \tag{2}$$

In Formula (2), 3D W-MSA refers to 3D Window-Based Multi-Head Attention, and 3D SW-MSA denotes 3D Shifted Windows Multi-Head Self-Attention [33].By integrating the operations of the KAN network, redundant operations are reduced while preserving the dependencies necessary for capturing adjacent features. This not only enriches contextual information and enhances the model's expressiveness but also slightly reduces computational complexity.

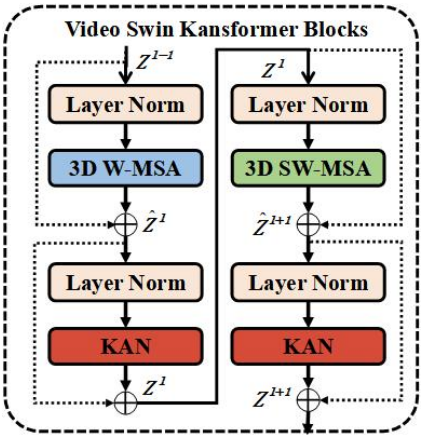


FIGURE 2. An illustration of two successive Video Swin Kansformer blocks

The Video Swin Kansformer is designed to effectively capture both spatial and temporal information in video data. Video processing is conceptually similar to time series analysis but extended to an additional spatial dimension. As shown in Equation (3), the video data consists of T frames, each with H×W×3 pixels. In the Video Swin Kansformer, each 3D patch of size 2×4×4×3 is treated as a token. This results in T/2×H/4×W/4 3D tokens, each containing 96-dimensional features. A linear embedding layer is then applied to project the features of each token to an arbitrary dimension denoted by C.

$$V \in \mathbb{R}^{T\times H\times W\times 3} \tag{3}$$

Consistent with existing techniques [34,35] and considering the characteristics of ultrasound data in this study, downsampling is not performed along the temporal dimension. This decision allows the model to maintain the hierarchical structure of the original Swin Transformer [36], which consists of four stages. Each stage performs 2× spatial downsampling in the patch merging layer. The patch merging layer concatenates features from each group of 2×2 spatially neighboring patches and applies a linear layer to project the concatenated features to half of their original dimensions..

C. Clinical Kansformer Blocks

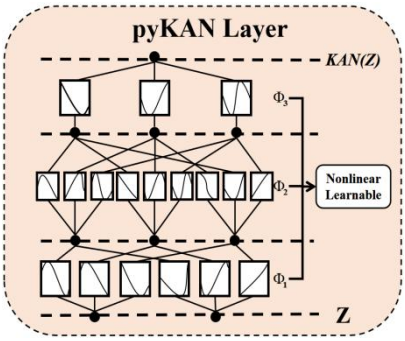


FIGURE 3. pykan LAYER structure diagram

Based on the study of KAN, this research developed the Clinical Kansformer Blocks to extract key features from clinical data. The structure of these blocks is shown in Figure 3. To accelerate the computational speed of the original pyKAN implementation [39], we utilized module code from the Efficient KAN project [38].

This architecture further optimizes the Transformer structure by incorporating a 1D Conv-KAN module and a Linear-KAN block. The design philosophy behind these modules is to replace traditional n-dimensional convolutional layers and linear multiplication operations with KAN operations [40]. In addition, the architecture leverages Multi-Head Attention to obtain attention distributions across different subspaces of the input sequence, allowing for a more comprehensive capture of potential pathological data associations. Specifically, the architecture follows these steps: first, it standardizes the input data using Layer Norm; then, the 1D Conv-KAN module extracts features, which are followed by a residual network after ReLU activation. Layer Norm is applied again to balance the data, and Multi-Head Attention is used to extract associations among pathological data. Finally, Layer Norm and the Linear-KAN block replace the fully connected layers commonly found in traditional neural networks and are followed by another residual network to complete feature processing. The Clinical Kansformer Blocks are defined as shown in Equation (4):

$$\begin{aligned} X^{t} &= Relu(1D\ Conv\text{-}kan(LN(X^{t-1})))\oplus X^{t-1} \\ X^{t+1} &= Attention(LN(X^{t}))\oplus X^{t} \\ X^{t+2} &= Linear\text{-}kan(LN(X^{t+1}))\oplus X^{t+1} \end{aligned} \tag{4}$$

D. Model Fusion for UThyroV-ST Swin Kansformer

Due to the significant differences in input token requirements between detection video data and clinical text data, the UTV-ST Swin Kansformer requires a specialized feature fusion approach. While four fusion methods have been explored in the literature [32] for combining one-dimensional temporal and video data, we adopted a staged slow fusion method, building on these successful approaches, to merge the data. This process is illustrated in Figure 4..

Figure 4(a) shows the first stage of data fusion, where the Video Swin Kansformer Blocks are trained with labeled video data to extract spatial and temporal features. Figure 4(b) depicts the second stage, where the Clinical Kansformer is trained using standardized text data. In Figure 4(c), the parameters of the Video Swin Kansformer Blocks are frozen to preserve the features they have learned, while a Classification Head integrates the feature representations from both Kansformer Blocks. This fusion process, executed within the Classification Head, is trained alongside the Clinical Kansformer, ensuring that video information serves as the primary feature and standardized clinical information acts as supplementary. This approach enables the model to achieve effective diagnostic results.

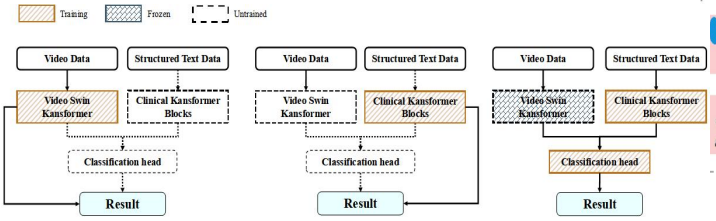


FIGURE 4. Segmented slow fusion method

IV. EXPERIMENT

A. Data Acquisition and Processing

We retrospectively collected patient data from the First Hospital of Shanxi Medical University, including individuals who underwent thyroid ultrasound examinations followed by surgical treatment between July 2022 and July 2023. The study was approved by the Ethics Committee of the First Hospital of Shanxi Medical University, and informed consent was waived.

During data collection, we strictly adhered to predefined inclusion and exclusion criteria. The inclusion criteria were: (1) patients who underwent total or subtotal thyroidectomy and neck lymph node dissection; (2) nodules pathologically confirmed as papillary thyroid carcinoma (PTC) through surgery; (3) patients who had routine ultrasound examinations within two weeks before surgery, with clear and complete original detection videos available. The exclusion criteria were: (1) patients who had received radiofrequency ablation, radiotherapy, or chemotherapy before surgery; (2) cases where ultrasound images of the target tumor were compromised by artifacts; (3) patients with other concurrent malignancies; (4) patients with a history of thyroid surgery.

According to the inclusion and exclusion criteria, a total of 346 patients were included in the study, with 352 thyroid nodules captured in 346 ultrasound videos. The dataset is comprehensive, containing not only the ultrasound videos but also detailed clinical information based on the postoperative pathological results for all 346 patients. The nodules were classified into three categories based on their metastatic status: Class I (no metastasis), Class II (metastasis to the central cervical lymph nodes), and Class III (metastasis to both central and lateral cervical lymph nodes). This rich dataset, combining real-time ultrasound imaging and clinical data, reflects a wide array of clinical scenarios and metastatic conditions, making it especially valuable for training and validating our model. The processing methods and detailed information of the clinical data can be found in Supplementary Table S1.

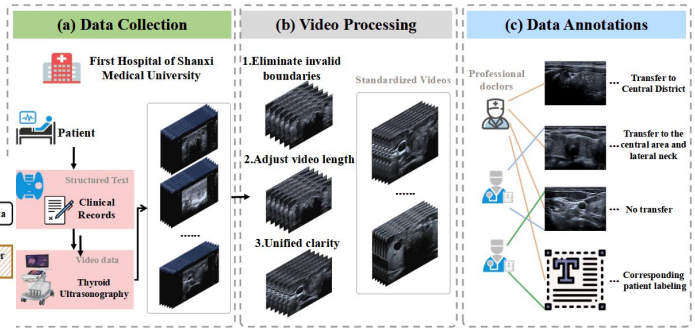


FIGURE 5. Workflow for building a dataset. (a) The data comes from patient thyroid ultrasound examination videos collected by the First Hospital of Shanxi Medical University. (b) The videos are standardized through processing software and systems to ensure that they are suitable for algorithm processing. (c) Expert doctors at the First Hospital of Shanxi Medical University carefully identify and manually annotate each video based on the patient's puncture biopsy results.

During data processing, several predictive features were initially classified as categorical variables or Boolean values. These data were then transformed to meet the model's requirements. Specifically, variables such as gender and whether the aspect ratio is greater than 1 were encoded in a binary (0-1) format. For instance, if the gender is "male," the value is assigned as 1, and if "female," the value is assigned as 0. This transformation converts categorical variables into binary format, making them easier for the algorithm to process. For continuous variables, such as size and age, feature scaling was applied to normalize the measurement scales of the various features. The normalization process can be expressed using the following formula (5):

$$X' = \frac{X - \mu}{\sigma} \tag{5}$$

where X is the original value, μ is the mean of the dataset, and σ is the standard deviation. This standardization ensures that the albumin levels have a

mean of 0 and a standard deviation of 1, enhancing algorithm performance and expediting model convergence.

For this study, we performed data preprocessing. First, irrelevant information from the videos, such as invalid borders and machine model details, was removed. Then, the video lengths were standardized. Finally, we adjusted the video clarity. After preprocessing, all video data were standardized to a length of 5 seconds. For longer videos, multiple 5-second clips were extracted, while shorter videos were looped to reach the 5-second duration. The videos were standardized to a resolution of 224×224 pixels, ensuring that the nodules in the detection images were clearly visible (Figure 5).

B. Experimental Settings

The experiments were conducted on a server platform with specific hardware configurations, using the PyTorch framework for algorithm development and model training. The hardware specifications included an i7-14700k CPU, 128 GB of RAM, and two RTX 4090 Ti GPUs (each equipped with 24 GB of VRAM), providing an efficient computational environment. The dataset was randomly split into training and testing sets in a 7:3 ratio, and five-fold cross-validation was employed during training. The standardized text information for each patient was processed using a consistent classification scheme.

The Clinical Kansformer utilized an embedding dimension of 24 and a window size of 4 to balance the model's receptive field. Each input data point was treated as an initial patch to ensure fine-grained analytical capability. The architecture consisted of a single module and employed four attention heads at all levels. A stochastic depth rate of 0.1 was introduced during training to improve regularization. The batch size was set to 500, and the Adam optimizer was used with an initial learning rate of 0.0005, decaying to 0.0001 after 100 epochs. The model was gradually optimized as it converged, and training continued until the validation loss reached its minimum.

The Video Swin Kansformer Blocks comprised four modules, with the number of attention heads in each module set to 4, 8, 16, and 32, respectively. A stochastic depth rate of 0.1 was employed during training, and the AdamW optimizer was used with a batch size of 8 and an initial learning rate of 0.001. The learning rate schedule was divided into two stages: initially, linear scaling of the learning rate using the LinearLR scheduler from epochs 0 to 200; thereafter, the learning rate was reduced to 90% of its previous value every 20 epochs, followed by validation. The epoch with the minimum validation loss was selected as the final model, ensuring that the model converged to an optimal solution.

For data fusion, a batch size of 8 was used, and the SGD optimization algorithm was chosen with a learning rate of 0.0001. Since the Video Swin Kansformer did not require further training, its batch size was set to 2000.

C. Classification Performance

To thoroughly evaluate the performance of different models on the clinically imbalanced thyroid ultrasound dataset, we employed the following standard evaluation metrics: accuracy, precision, recall, and F1-score, calculated using Equation (6).

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+FP+FN+TN} \\ Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+TN} \\ F_1 &= 2\frac{Precision \times Recall}{Precision + Recall} \end{aligned} \tag{6}$$

True positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are defined as follows: TP and TN represent samples correctly predicted as positive and negative, respectively, while FP and FN represent samples incorrectly predicted as positive and negative, respectively. In addition, we assessed the models using the average AUC [41], derived from the classification confidence.

The Clinical Kansformer demonstrated outstanding performance in invasiveness classification, with both accuracy and average AUC surpassing other benchmarks (Table 1). These results indicate that the Clinical Kansformer within the UTV-ST framework set a new state-of-the-art performance standard in classifying standardized clinical information.

TABLE 1
COMPARISON OF CLASSIFICATION RESULTS OF DIFFERENT CLASSIFIERS ON STANDARDIZED DATA

Method	Ave. Acc.	Precision	Recall	F1 score	Ave. AUC
Logistic Regression	0.692	(0.85,0.65,0.50)	(0.83,0.63,0.45)	(0.84,0.64,0.47)	0.675
Decision tree	0.716	(0.87,0.68,0.53)	(0.85,0.67,0.48)	(0.86,0.67,0.50)	0.700
Naive Bayes	0.732	(0.88,0.70,0.55)	(0.87,0.69,0.50)	(0.87,0.69,0.52)	0.715
XGBoost	0.765	(0.90,0.73,0.58)	(0.89,0.72,0.54)	(0.89,0.72,0.56)	0.740
SVM	0.773	(0.92,0.75,0.60)	(0.91,0.74,0.56)	(0.91,0.75,0.58)	0.755
Random Forest	0.790	(0.94,0.77,0.63)	(0.93,0.76,0.60)	(0.93,0.76,0.61)	0.770
MLP	0.802	(0.95,0.79,0.65)	(0.94,0.78,0.62)	(0.94,0.78,0.63)	0.785
Clinical Kansformer	0.823	(0.97,0.82,0.68)	(0.96,0.81,0.65)	(0.96,0.81,0.66)	0.837

The proposed UTV-ST Swin Kansformer was compared with other benchmark models, and the results are shown in Table 2. To ensure fairness, all models used the same training, validation, and testing datasets, and the configurations of the benchmark models strictly followed the specifications in the original literature. In all multimodal fusion processes, an MLP was consistently used as the classification method for standardized clinical information. Table 3 shows that the proposed model

outperforms all benchmark models in terms of accuracy, AUC, and F1-score. These results highlight the model's excellent performance in classifying thyroid nodule invasiveness.

Comparison among models reveals that our model surpasses most benchmark models in AP and AP50 metrics, notably achieving 94.2% in AP50. Additionally, the FLOPs and number of parameters are relatively reasonable, indicating that the model strikes a good balance between performance and computational efficiency.

TABLE 2
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON MULTIMODAL JOINT ULTRASONIC TESTING DATASET

Method	Pretrain	AP^{val}	AP_{ω}^{val}	Views	FLOPs	Param
TimeSformer [42]	ImageNet-21K	81.8	91.1	1×3	1703	121.4
SlowFast R101+NL [43]	—	78.2	92.3	1×3	234	59.9
X3D-XXL [44]	—	76.5	91.7	1×1	129	40.2
ViViT-L/16x2 [45]	ImageNet-21K	79.1	92.5	—	903	352.1
MViT-B, 32×3 [46]	—	81.0	92.8	1×3	455	36.6
MViT-B, 64×3 [46]	—	80.3	93.4	1×3	236	236
Video swin Transformer [33]	ImageNet-1K	80.8	93.8	1×3	321	88.8
SwinVid[47]	ImageNet-1K	81.1	93.7	1×3	372	186
YOLOv11x[48]	ImageNet-1K	72.1	86.2	—	194.9	56.9
DuST[32]	ImageNet-1K	80.9	92.5	1×3	257	43.4
Our	ImageNet-1K	82.1	94.2	1×3	353	75.4

D. Ablation Study

Table 3 presents the contributions of various components integrated into the UTV-ST Swin Kansformer network for detection on the standardized text-video thyroid dataset. During testing, the fusion methods from the Dual Swin Transformer paper were adopted. The baseline model used Swin Transformer-B as its foundation, and when calculating accuracy, the view setting was 1×3. In the absence of the Clinical Kansformer, an MLP was used as a substitute.

TABLE 3
:PERFORMANCE OF DIFFERENT COMPONENTS IN THE UTV-ST SWIN KANSFORMER NETWORK

Method	Video Swin Kansformer		Clinical Kansformer	AP^{val}	AP_{50}^{val}
	KAN	Video Swin Transformer			
1				75.4	83.7
2	√			77.2	85.0
3			√	78.6	87.1
4		√	√	80.3	89.5
5	√	√		81.2	92.3
6	√	√	√	82.1	94.2

The results in Table 3 indicate that when the UTV-ST Swin Kansformer model is unimodal (Model 5), the accuracy reaches 81.2%, and the mean Average AP_{50}^{val} is 92.3%, demonstrating good performance. Under multimodal fusion (Model 6), the results are optimal, with the AP reaching 82.1% and AP_{50}^{val} as high as 94.2%. These findings provide important experimental evidence for the optimization of the UTV-ST Swin Kansformer architecture, suggesting that a reasonable combination of modules can significantly enhance the model's ability to process detailed features.

E. Visualization of the detection results

We evaluated the multimodal ultrasound dataset using the UTV-ST Swin Kansformer model, and the results are shown in Figure 6. The frames with the highest confidence in the video were displayed using Python's Matplotlib library. From the figure, it can be observed that the confidence for Class 2 images was relatively low, but the overall classification accuracy was high, particularly showing advantages in classifying non-invasive nodules. The UTV-ST Swin Kansformer performs excellently in detecting the invasiveness of thyroid nodules, capable of accurately distinguishing nodule invasiveness and effectively extracting key lesion features from combined visual and textual data.

To gain deeper insights into the key factors influencing the invasiveness of thyroid nodules, we conducted a comprehensive analysis of the Clinical Kansformer Blocks and identified the most influential factors contributing significantly to nodule invasiveness. This analysis also aids physicians in assessing the necessity of considering certain factors when evaluating patients (Supplementary Figure S1). As shown in the figure, age, adjacency or invasion (adjacent to or infringing upon), and suspicious lymph nodes account for a substantial proportion.

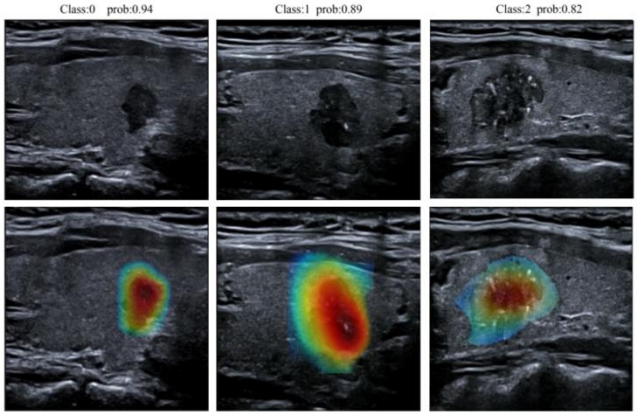


FIGURE 6. Multimodal ultrasound evaluation results

F. Limitations and shortcomings

Despite the promising results, there are some limitations that should be addressed for broader clinical applicability. First, the variability in data quality, including differences in imaging devices and patient demographics, could affect the model's generalizability to real-world clinical scenarios. While the datasets used in this study were extensive, they may not fully capture the diversity of conditions encountered in diverse clinical environments. Expanding the dataset to include more varied patient profiles and imaging conditions, along with employing

data augmentation techniques, could improve the model's robustness and generalizability.

Another limitation is the computational complexity of the model, which presents challenges for deployment in resource-constrained clinical settings. The large number of parameters and high memory requirements could make it difficult to apply the model in real-time applications. To address this, future work will focus on model optimization techniques, such as pruning, quantization, or knowledge distillation, to reduce its size and improve efficiency without sacrificing accuracy. These improvements will make the model more feasible for use in clinical environments with limited computational resources.

V. Conclusion

In this study, we proposed the UTV-ST Swin Kansformer model, specifically designed to assess the invasiveness of thyroid nodules by integrating multimodal data, including ultrasound videos and standardized clinical information. The main advantage of this model is its ability to combine the spatial and temporal information of ultrasound videos with key clinical data, providing a comprehensive understanding of nodule invasiveness. This integration reduces reliance on operator expertise, which is a common limitation in ultrasound diagnosis.

Although the UTV-ST Swin Kansformer model demonstrates excellent performance, it does have some limitations. First, it relies on a specialized dataset, and the availability of public data that fits the model ' s requirements is relatively limited. This constraint affects its generalizability across different populations and imaging devices. Second, the model ' s performance can still be improved, particularly in extremely complex or low-quality imaging conditions. Additionally, the computational complexity of the model is high, which may hinder its real-time application in resource-limited clinical settings. Future research should focus on enhancing both the robustness and efficiency of the model. Increasing the size and diversity of the dataset, especially by including more complex cases and data from various devices, will help improve the model ' s generalizability. Simultaneously, optimizing the computational efficiency is crucial to ensure its applicability in resource-constrained environments. There is significant potential in integrating real-time diagnostic capabilities, and exploring the application of this model in other medical imaging fields could yield promising results.

ACKNOWLEDGMENT

This study sincerely thanks Tiangong University for providing valuable technical and equipment support in this research.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request. Additionally, the study utilizes publicly available datasets, which can be accessed through references.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Funding

This study was supported by Key Research and Development Program of Science and Technology, Department of Shanxi Province (201903D321190) and Ministry of Human and Social Affairs Program for the selection of Science and Technology activities for Returned Overseas Scholar (2016-366).

REFERENCES

[1]Zhang J, Zhang F, Zhao C, Xu Q, Liang C, Yang Y, et al. Dysbiosis of the gut microbiome is associated with thyroid cancer and thyroid nodules and correlated with clinical index of thyroid function. Endocrine. 2019;64(3):564–74. https://doi.org/10.1007/s12020-018-1831-x.

[2] Kim S M, Pereira J A, Lopes Jr V, et al. Practical active control of cavity noise using loop shaping: Two case studies[J]. Applied Acoustics, 2017, 121: 65-73.

[3] Ho A S, Sarti E E, Jain K S, et al. Malignancy rate in thyroid nodules classified as Bethesda category III (AUS/FLUS)[J]. Thyroid, 2014, 24(5): 832-839.

[4] Kuru B, Atmaca A, Kefeli M. Malignancy rate associated with Bethesda category III (AUS/FLUS) with and without repeat fine needle aspiration biopsy[J]. Diagnostic cytopathology, 2016, 44(5): 394-398.

[5] Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. J Am Coll Radiol. 2017;14(5):587–95. https://doi.org/10.1016/j.jacr. 2017.01.046.

[6] Park JY, Lee HJ, Jang HW, Kim HK, Yi JH, Lee W, et al. A proposal for a thyroid imaging reporting and data system for ultrasound features of thyroid carcinoma. Thyroid. 2009;19(11):1257–https://doi.org/10.1089/thy.2008.0021.

[7] Hoang JK, Middleton WD, Farjat AE, Langer JE, Reading CC, Teefey SA, et al. Reduction in thyroid nodule biopsies and improved accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. Radiology. 2018;287(1):185–93. https://doi.org/10.1148/radiol.2018172572.

[8] Yoon JH, Han K, Kim EK, Moon HJ, Kwak JY. Diagnosis and management of small thyroid nodules: a comparative study with six guidelines for thyroid nodules. Radiology. 2017;283(2):560–9. https://doi.org/10.1148/radiol.2016160641.

[9] Li F, Sun W, Liu L, Meng Z, Su J. The application value of CDFI and SMI combined with serological markers in distinguishing benign and malignant thyroid nodules. Clin Transl Oncol. 2022;24(11):2200–9. https://doi.org/10.1007/ s12094-022-02880-1.

[10]Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. Thyroid. 2010;20(2):167–72. https://doi.org/10.1089/thy. 2008.0354.

[11]Zhan J, Zhang LH, Yu Q, Li CL, Chen Y, Wang WP, et al. Prediction of cervical ymph node metastasis with contrast-enhanced ultrasound and association between presence of BRAFV600E and extrathyroidal extension in papillary thyroid carcinoma. Ther Adv Med Oncol (2020) 12:1758835920942367. doi: 10.1177/1758835920942367

[12]Li WH, Yu WY, Du JR, Teng DK, Lin YQ, Sui GQ, et al. Nomogram prediction for cervical lymph node metastasis in multifocal papillary thyroid microcarcinoma. Front Endocrinol (Lausanne) (2023) 14:1140360. doi: 10.3389/fendo.2023.1140360

[13]Chang L, Zhang Y, Zhu J, Hu L, Wang X, Zhang H, et al. An integrated nomogram combining deep learning, clinical characteristics and ultrasound features for predicting central lymph node metastasis in papillary thyroid cancer: A multicenter study. Front Endocrinol (Lausanne) (2023) 14:964074. doi: 10.3389/fendo.2023.964074
[14]Liu Q, Li Y, Hao Y, et al. Multi-modal ultrasound multistage classification of PTC cervical lymph node metastasis via DualSwinThyroid[J]. Frontiers in Oncology, 2024, 14: 1349388.
[15]Inan N G, Kocadağlı O, Yıldırım D, et al. Multi-class classification of thyroid nodules from automatic segmented ultrasound images: Hybrid ResNet based UNet convolutional neural network approach[J]. Computer Methods and Programs in Biomedicine, 2024, 243: 107921.
[16]He H, Zhu J, Ye Z, et al. Using multimodal ultrasound including full-time-series contrast-enhanced ultrasound cines for identifying the nature of thyroid nodules[J]. Frontiers in Oncology, 2024, 14: 1340847.
[17]Du H, Chen F, Li H, et al. Deep-learning radiomics based on ultrasound can objectively evaluate thyroid nodules and assist in improving the diagnostic level of ultrasound physicians[J]. Quantitative Imaging in Medicine and Surgery, 2024, 14(8): 5932.
[18]Rho M, Chun S H, Lee E, et al. Diagnosis of thyroid micronodules on ultrasound using a deep convolutional neural network[J]. Scientific reports, 2023, 13(1): 7231.
[19]Chen C, Jiang Y, Yao J, et al. Deep learning to assist composition classification and thyroid solid nodule diagnosis: a multicenter diagnostic study[J]. European Radiology, 2024, 34(4): 2323-2333..
[20]Zhang N, Liu J, Jin Y, et al. An adaptive multi-modal hybrid model for classifying thyroid nodules by combining ultrasound and infrared thermal images[J]. BMC bioinformatics, 2023, 24(1): 315.
[21]]Liu Q, Li Y, Hao Y, et al. Multi-modal ultrasound multistage classification of PTC cervical lymph node metastasis via DualSwinThyroid[J]. Frontiers in Oncology, 2024, 14: 1349388.
[22]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017
[23]Guanyu Chen, Peng Jiao, Qing Hu, Linjie Xiao, and Zijian Ye. Swinstfm: Remote sensing spatiotemporal fusion using swin transformer. IEEE Transactions on Geoscience and Remote Sensing, 60:1–18, 2022. 2
[24]Guanyu Chen, Tianyi Shi, Baoxing Xie, Zhicheng Zhao, Zhu Meng, Yadong Huang, and Jin Dong. Swindae: Electrocar diogram quality assessment using 1d swin transformer and denoising autoencoder. IEEE Journal of Biomedical and Health Informatics, 27(12):5779–5790, 2023.
[25]Ricard Lado-Roige and Marco A P ´ erez. Stb-vmm: Swin ´ transformer based video motion magnification. KnowledgeBased Systems, 269:110493, 2023. 2
[26]Ze hui Li, Akashaditya Das, William A V Beardall, Yiren Zhao, and Guy-Bart Stan. Genomic interpreter: A hierarchical genomic deep neural network with 1d shifted window transformer, 2023. 2
[27]Soren Richard Stahlschmidt, Benjamin Ulfenborg, and Jane ¨ Synnergren. Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics, 23(2):bbab569, 2022. 2
[28]Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014. 2
[29]Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2
[30]Maria Shoukat, Khubaib Ahmad, Naina Said, Nasir Ahmad, Mohammed Hassanuzaman, and Kashif Ahmad. A late fusion framework with multiple optimization methods for media interestingness. arXiv preprint arXiv:2207.04762, 2022.2
[31]Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems, 35:32897–32912, 2022. 2
[32]Shi L, Chen Y, Liu M, et al. DuST: Dual Swin Transformer for Multi-modal Video and Time-Series Modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 4537-4546.
[33]Liu Z, Ning J, Cao Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 3202-3211..
[34]Qiu, Z., Yao, T., and Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In proceedings of the IEEE International Conference on Computer Vision, pages 5533–5541.
[35]Xie, S., Sun, C., Huang, J., Tu, Z., and Murphy, K. (2018). Rethinking spatiotemporal featurelearning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 305–321.
[36]Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprinarXiv:2103.14030.
[37]Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljaci ˇ c, ´ T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-arnold networks," arXiv preprint arXiv:2404.19756, 2024.
[38]Efficient KAN. Available online: https://github.com/Blealtan/efficient-kan (accessed on 06 July 2024).
[39]PyKAN. Available online: https://github.com/KindXiaoming/pykan (accessed on 06 July 2024).
[40]ConvKAN. Available online: https://github.com/StarostinV/convkan (accessed on 06 July 2024).
[41]Nancy A Obuchowski and Jennifer A Bullen. Receiver operating characteristic (roc) curves: review of methods with applications in diagnostic medicine. Physics in Medicine & Biology, 63(7):07TR01, 2018. 6
[42]Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095.
[43]Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6202–6211.
[44]Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 203–213.
[45]Arnab, A., Dehghani, M., Heigold, G., Sun, C., Luci ˇ c, M., and Schmid, C. (2021). Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691.
[46]Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., and Feichtenhofer, C. (2021b). Multiscale vision transformers. arXiv preprint arXiv:2104.11227.

Yufang Zhao is an attending physician and master's degree holder. He is a Youth Member of the Ultrasound Medicine Branch of the China Ethnic Health Association, a member of the Superficial Organs and Musculoskeletal Ultrasound Committee of the China Medical Education Association, a Youth Member of the 2nd Breast Cancer Committee of the Shanxi Anti-Cancer Association, and a member of the Ultrasound Branch of the Shanxi Geriatrics Society.

Dr. Zhao specializes in ultrasound diagnosis of the breast, thyroid, abdomen, and male reproductive system, as well as in ultrasound-guided interventions such as tumor biopsy, percutaneous catheter drainage, and cyst sclerotherapy.

Yue Li was born in Shanxi, China, in 1995. He is currently a Ph.D. candidate at the School of Electronics and Information Engineering, Tianjin Polytechnic University.

His research focuses on medical artificial intelligence and computer vision, with a strong emphasis on deep learning techniques. Li has published over five SCI-indexed papers related to deep learning, contributing to advancements in the application of AI in the medical field. His

work aims to leverage deep learning to improve diagnostic accuracy and treatment outcomes in various medical applications.

Yanjing Zhang is an attending physician and master's degree holder. She is a Youth Member of the Musculoskeletal Branch of the Chinese Ultrasound Medical Engineering Society, a Youth Member of the Musculoskeletal and Superficial Ultrasound Committee of the Chinese Research Hospital Association, a member of the Shanxi Ultrasound Medical Engineering Society, and a member of the Shanxi Geriatrics Ultrasound Society.
Dr. Zhang specializes in ultrasound diagnosis and treatment, particularly in abdominal, superficial, and peripheral vascular diseases. She is especially skilled in diagnosing and providing interventional treatment for common musculoskeletal conditions, including muscle, tendon, joint, and nerve disorders.

Xiaohui Yan is a physician and master's degree holder. She has participated in 1 National Natural Science Foundation project and 2 provincial and ministerial research projects. Dr. Yan has published 5 papers in SCI journals and domestic core journals.

She received the 2022 Excellent Paper Award from the Ultrasound Professional Committee of the Chinese Women Physicians Association. Her research focuses on interventional ultrasound as well as ultrasound diagnosis of the breast, thyroid, and abdomen.

Guolin Yin is a physician and master's degree holder. He specializes in ultrasound diagnosis of the abdomen, thyroid, breast, and male reproductive system, as well as in ultrasound-guided interventions such as tumor biopsy and percutaneous catheter drainage. Dr. Yin has published 5 papers in SCI journals and domestic core journals.

Liping Liu received her doctorate in Clinical Medicine from the PLA General Hospital in 2006. She is currently the Director, Professor, and Doctoral Supervisor of the Ultrasound Intervention Department at the First Hospital of Shanxi Medical University. Her primary research focus is in ultrasound medicine, with particular interests in ultrasound imaging, contrast agents, and interventional ultrasound technology, especially in the diagnosis and treatment of tumors.
In 2006, Dr. Liu was a visiting scholar at the Ultrasound Medicine Center of Thomas Jefferson University Hospital in the United States. In 2015, she also served as a senior visiting scholar at the Department of Radiology at the University of Southern California Hospital. Dr. Liu has led 2 National Natural Science Foundation projects and 13 provincial and ministerial research projects. The projects she presided over won 2 second prizes for Scientific and Technological Progress in Shanxi Province.