IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 419-431
Journal homepage: http://iieta.org/journals/ts

Using Deep Learning Technique, Multi-Task Identification of Abnormal Positioned Tooth in Maxillary Sinus Through a Method that Combines 2D CNN and Transformer Technology

Vimala R.[1*], D. M. D. Preethi [2]

[1] Department of Computer Science and Engineering, Government College of Engineering, Bodinayakanur 6225582, India
[2] Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul 624622, India

Corresponding Author Email: vimalaleela09@gmail.com

https://doi.org/10.18280/ts.430129

Received: 7 January 2026
Revised: 15 February 2026
Accepted: 25 February 2026
Available online: 28 February 2026

Keywords:
abnormally positioned tooth, CNN, maxillary sinus, 2D Hybrid CNN-Transformer, Dynamic Spatial Attention

ABSTRACT

The detection and treatment of abnormally positioned tooth within the maxillary sinus present significant clinical challenges, resulting in complications such as Odontogenic factors are reported amounting to approximately 10–40% of maxillary sinusitis cases. These challenges are intensified by limited access to the sinus cavity, close proximity to vital anatomical structures, and the potential for postoperative issues such as sinus membrane damage. Therefore, accurate detection and localization are essential for planning safe and effective surgical intervention. This research proposes a multi-task model built using deep learning techniques for identifying and localizing abnormally positioned tooth within the maxillary sinus, utilizing a 2D Hybrid CNN–Transformer architecture enhanced with Dynamic Spatial Attention (DSA). Convolutional layers are employed for extracting local spatial features, while Transformer modules learn global contextual information. The integration of spatial attention mechanisms further improves detection precision and anatomical awareness. Trained on a dataset of 300 CBCT scans, the framework achieved 94% classification accuracy, 87.5% precision, 93.3% recall, and 88.7% mean Average Precision (mAP). Comparative benchmarking against ResNet50–YOLOv5 and U-Net–CNN demonstrates improved detection robustness and enhanced generalization capability. The integration of attention-guided feature refinement and optimized post-processing establishes a computationally efficient and anatomically consistent solution for automated maxillary sinus pathology assessment, supporting AI-assisted diagnosis and surgical planning in maxillofacial radiology.

1. INTRODUCTION

Dentistry represents a specialized branch of medicine dedicated to the evaluation, prevention and treatment of conditions affecting the oral and maxillofacial regions, which are integral to overall systemic health. Among the various clinical challenges encountered in dental practice, the presence of displaced teeth within the maxillary sinus constitutes a complex diagnostic and therapeutic concern. Ongoing advancements in dental research have highlighted the necessity for accurate assessment and carefully planned management strategies to address such anomalies effectively [1-3]. The anatomical association in the region posterior tooth in the maxillary region and the inferior boundary of the maxillary sinus cavity is of considerable clinical importance. In the upper jaw, the minimal osseous barrier separating the sinus cavity from the apices of posterior tooth increases the likelihood of sinus involvement during dental extractions, where displacement of roots into the sinus may occur as a recognized complication [4]. Consequently, precise evaluation of this region is essential to minimize procedural risks and ensure safe clinical outcomes.

A detailed understanding of both normal anatomy and structural variations involving the maxillary sinus is critical for precise diagnosis and proper treatment planning. Anatomically, the maxillary sinus is pyramid-shaped, featuring its base forming the lateral wall of the nasal cavity and its apex directed toward the zygomatic bone. It generally exhibits dimensions of approximately 35 mm in height, 25 mm in width and 35–45 mm in length, with an average volume near 15 mL, though these parameters vary according to age and individual anatomical differences [5]. Variations such as sinus septa, accessory ostia, sinus hypoplasia and atypical dental root morphology can alter sinus physiology and complicate surgical procedures [6, 7]. These anatomical intricacies emphasize the importance of accurate imaging and detailed structural analysis.

From a clinical perspective, teeth displaced into the sinus cavity may result in complications including chronic sinus inflammation, nasal obstruction, facial discomfort, dentigerous cyst development and oroantral communication. Effective management typically requires coordinated collaboration among oral and maxillofacial surgeons, radiologists, and restorative specialists. Therapeutic decisions must be tailored to the specific anatomical position and pathological features of the anomaly [8]. When the displaced tooth is located near the sinus floor, surgical approaches such as transalveolar or transantral techniques may be employed to

reduce operative risks and preserve surrounding structures [9]. Advances in surgical technologies have further contributed to improved precision and enhanced patient outcomes [10]. Given the structural complexity of the maxillary sinus and the potential for serious complications, accurate detection and localization are crucial. Nevertheless, manual interpretation of radiographic data can be labor-intensive and prone to variability among clinicians. The presence of overlapping bony structures and heterogeneous image intensities within the sinus region further complicates consistent evaluation and reduces diagnostic reliability.

In response to these limitations deep learning techniques, are increasingly incorporated into dental imaging workflows. Deep learning–driven identification systems enable accurate localization, spatial orientation assessment, and analysis of the relationship between displaced tooth and adjacent anatomical landmarks, thereby supporting informed clinical decision-making [11]. Automated detection and segmentation frameworks not only enhance diagnostic consistency but also facilitate interdisciplinary treatment planning. This work contributes to the growing literature through the development of an automated identification and detection framework tailored for evaluating ectopic tooth within the maxillary sinus while evaluating their positional relationships with important anatomical structures, ultimately improving diagnostic precision and therapeutic planning.

2. RELATED WORK

This section describes a comprehensive clinical and computational strategy for evaluating patients suspected of having ectopic or displaced tooth within the maxillary sinus. An accurate diagnostic process necessitates the integration of detailed medical and dental history assessment with advanced radiographic imaging techniques. The evaluation begins with systematic documentation of patient symptoms, as reported in the study [12], along with an analysis of predisposing factors such as previous trauma, surgical procedures and developmental irregularities that may contribute to abnormal tooth positioning. The clinical examination encompasses both extra oral and intraoral assessments. Extra oral evaluation involves examining facial symmetry, localized tenderness and nasal discharge, whereas intraoral inspection focuses on detecting missing or impacted tooth, fistulous tracts, swelling, and abnormal mobility [13]. These observations serve as essential preliminary indicators that guide further radiological investigation. CBCT has evolved into extensively utilized imaging technique in maxillofacial diagnostics attributable to its capability to produce high-resolution two dimensional images. This imaging technique facilitates accurate assessment of the spatial orientation of displaced tooth relative to adjacent anatomical structures within the sinus cavity, thereby aiding in precise diagnosis and surgical decision-making [14]. Compared to conventional CT, CBCT provides advantages such as lower radiation dosage, faster image acquisition and comprehensive multiplanar reconstruction including coronal, axial, sagittal and multi-planar views.

Although CBCT significantly enhances visualization, manual evaluation of CBCT scans requires significant effort and may lead to inter-observer variability, particularly in anatomically intricate regions. The occurrence of overlapping bony structures, heterogeneous intensity distributions and poorly defined boundaries complicates consistent interpretation. As a result automated image analysis approaches have become increasingly important. Deep learning frameworks, especially Fully Convolutional Network (FCN) and U-Net-based architectures, have shown strong performance in medical image segmentation applications [15]. These convolutional models effectively learn hierarchical local features and maintain structural information through encoder–decoder designs with skip connections. Nevertheless, their restricted receptive fields restrict their potential to represent long-range contextual dependencies, which are essential in anatomically complex environments like the paranasal sinus region. To enhance multiscale contextual understanding, architectures such as PSP Net introduced pyramid pooling strategies for improved global feature aggregation. While these methods strengthened contextual representation, they still faced challenges in modeling extensive spatial dependencies, often resulting in segmentation errors in complex or low-contrast regions [16].

To overcome these shortcomings, we introduce a Hybrid CNN–Transformer architecture incorporating Dynamic Spatial Attention (DSA) [17]. This framework combines convolutional operations for effective local Feature representation learning via transformer self-attention modules enables modeling of global context and inter-pixel relationships across the full image. In contrast to conventional CNN-based models that emphasize localized feature learning, the integration of self-attention facilitates modeling of long-range spatial interactions and complex anatomical relationships. This hybrid configuration improves discrimination of overlapping structures and enhances localization accuracy in CBCT images of the maxillary sinus.

3. PROPOSED METHOD

This section describes the methodology for designing and implementing a deep learning system to identify an abnormally positioned tooth in the maxillary sinus. The task focuses on detecting such tooth in both the left and right maxillary sinus. The workflow begins with data collection, during which a dataset of axial CBCT images is acquired. Subsequently, annotation is performed by marking the images with precise indicators of abnormal tooth locations. Once the dataset is prepared, the model selection phase involves choosing an appropriate deep learning architecture. The selected architecture is trained using the labeled dataset. Following training, the model performance is analyzed based on selected criteria including recall, F1-score, and precision. During post-processing, the model outputs are refined and visualization techniques are applied to improve interpretability. Finally, testing is performed on previously unseen data to assess robustness and generalization capability.

3.1 Dataset and annotation

The dataset used for training and evaluation comprises 300 annotated axial CBCT images of the maxillary sinus, sourced from Cumbum United Scan. All imaging data were retrospectively collected from a clinical repository and anonymized prior to analysis. Expert radiologists manually annotated the images to identify abnormally positioned tooth within the maxillary sinus region. Each image was annotated with a three-class categorical label representing normal condition, left maxillary sinus abnormality, or right maxillary

sinus abnormality, encoded using one-hot vectors. For abnormal cases, precise localization was provided using bounding box annotations specified using the coordinates $\{X_{min}, Y_{min}, X_{max}, Y_{max}\}$. Out of the total dataset, 240 images were allocated for training, while 60 images were reserved for independent testing. Additionally, annotations distinguished the anatomical laterality of the abnormal tooth, specifying its presence in either the left or right maxillary sinus. The annotation protocol defined three distinct classes to facilitate structured learning and clinical interpretability in Table 1. Ground-truth labels were established through careful manual delineation of the tooth structures within the sinus cavity to ensure accurate spatial representation for both classification and detection tasks.

Table 1. Classes name and description

No.	Class Name	Class Description
1	Maxillary sinus	Maxillary sinus without an abnormally positioned tooth
2	Tooth in left maxillary sinus	An abnormally positioned tooth in left maxillary sinus
3	Tooth in right maxillary sinus	An abnormally positioned tooth in right maxillary sinus

3.2 Preprocessing

Preprocessing is applied to ensure data uniformity and enhance training stability. All Cone Beam CT images are adjusted to a resolution of to a dimension of standardized 128 × 128 pixels resolution to provide consistent input dimensions to the network. During resizing, the original aspect ratio is preserved to avoid geometric distortion, and padding is applied when necessary to achieve the required dimensions. Figure 1 shows a comparison of image resolution (height and width) before and after resizing, demonstrating the reduction in spatial dimensions after preprocessing. To further stabilize training, the pixel intensities are normalized to the interval [0,1] through division by 255, thereby minimizing the influence of intensity variability and improving numerical consistency [18]. For the classification task, ground-truth labels are converted into one-hot encoded vectors to support multi-class learning. Images without abnormalities are encoded as [1,0,0], those containing an abnormally positioned tooth located in the cavity of left maxillary sinus as [0,1,0], and those in the cavity of right maxillary sinus as [0,0,1]. This representation enables the model to output probabilistic class predictions. For the localization task, bounding box are normalized by dividing horizontal coordinates by the image width (W) and vertical coordinates by the image height (H), ensuring that all coordinate values are mapped to the range [0,1]. Such normalization provides consistent spatial scaling and improves the framework capability to generalize to images of varying original sizes.

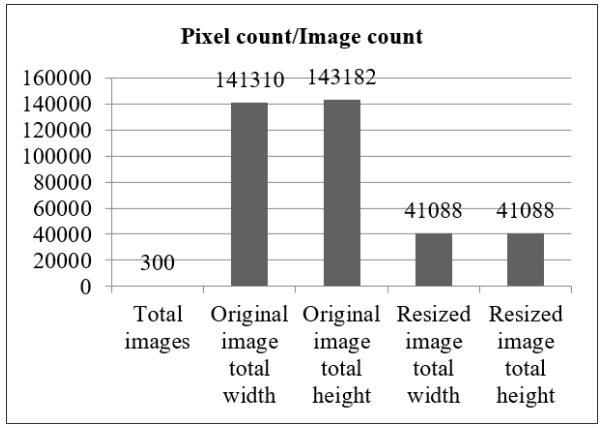


Figure 1. Image resolution height and width comparison of prior to and following resizing

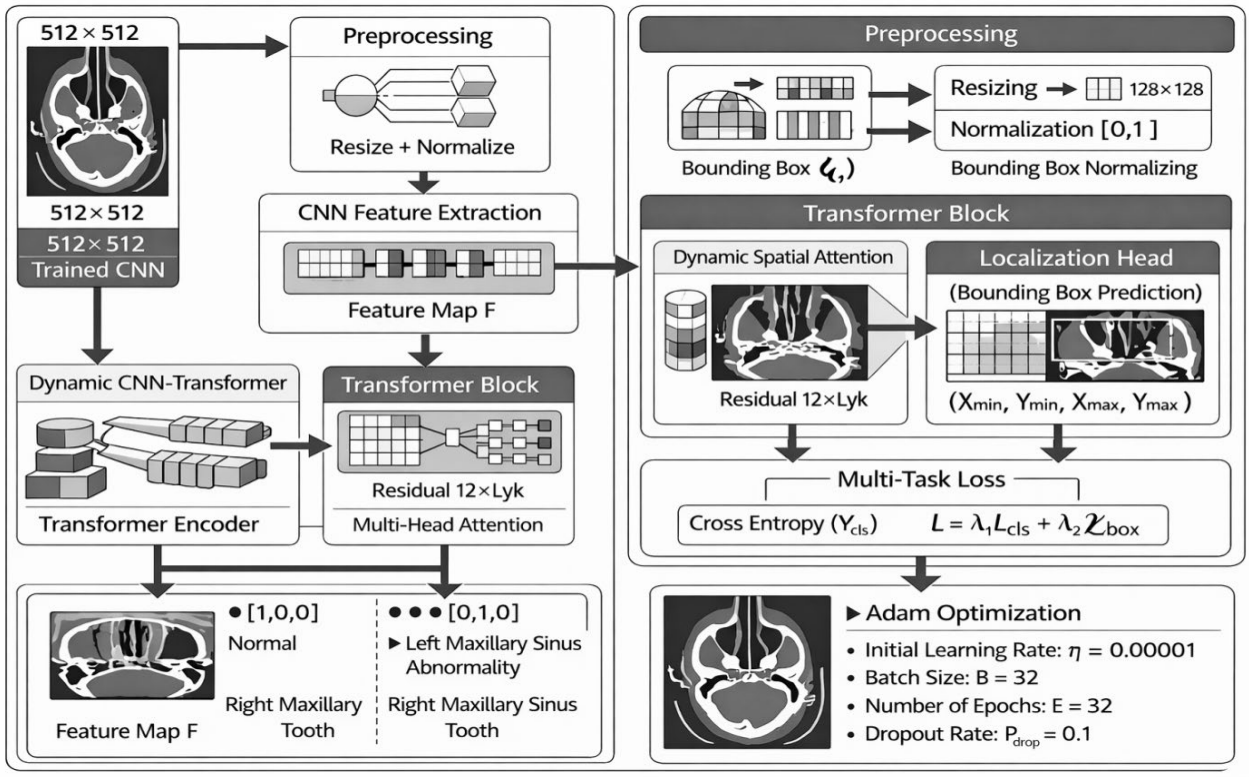


Figure 2. The 2D Hybrid CNN-Transformer with Dynamic Spatial Attention (DSA) architecture

3.3 Model architecture

The proposed framework, depicted in Figure 2, is built upon a 2D Hybrid CNN–Transformer architecture enhanced with a DSA mechanism within a multi-task learning setting. The model is specifically designed to combine detailed local feature learning combined with global contextual modeling for accurate detection and localization of abnormally positioned tooth in CBCT images of the maxillary sinus. The input CBCT image, resized to 128 × 128 pixels, is first processed by a CNN-based neural network backbone to obtain features learned at different hierarchical levels. These convolutional layers effectively capture local anatomical characteristics such as structural boundaries, shape variations, and texture patterns associated with sinus and dental regions [19]. To further improve discriminative capability, a DSA module is applied to the extracted feature maps. This module generates adaptive attention weights that emphasize clinically relevant regions while reducing the influence of non-informative background areas, thereby enhancing sensitivity to abnormal tooth structures. The processed features are subsequently transformed and passed into a Transformer encoder, where the self-attention based mechanism models long-distance spatial relationships and global contextual feature interactions across the entire image. Through the incorporation of local inductive bias of CNNs coupled along with the global representation the capability of Transformers, the hybrid architecture produces a more robust and comprehensive feature embedding.

The resulting feature representation is directed to two dedicated prediction branches under a unified multi-task framework. The classification branch determines the anatomical status of the maxillary sinus using three-class prediction (normal, left-sided abnormality and right-sided abnormality), whereas the localization branch estimates the corresponding bounding box. The model employs Categorical Cross-Entropy (CCE) loss to optimize three-class classification task, while Mean Squared Error (MSE) loss for bounding box prediction. This work focuses on a weighted sum pertaining to the loss functions. Model parameters are iteratively updated via backpropagation to minimize this joint loss. During training, the network simultaneously learns from classification annotations and bounding box labels, enabling concurrent anomaly detection and spatial localization shown in Training Alogrithm1. In the testing phase, the trained Hybrid CNN–Transformer with DSA processes unseen CBCT images and produces a three-class prediction (normal, left-sided abnormality, or right-sided abnormality) along with the associated bounding box coordinates, as illustrated in Figure 3. The framework is evaluated through distinct training and testing stages to assess both classification accuracy and localization effectiveness for abnormalities within the left and right maxillary sinus regions.

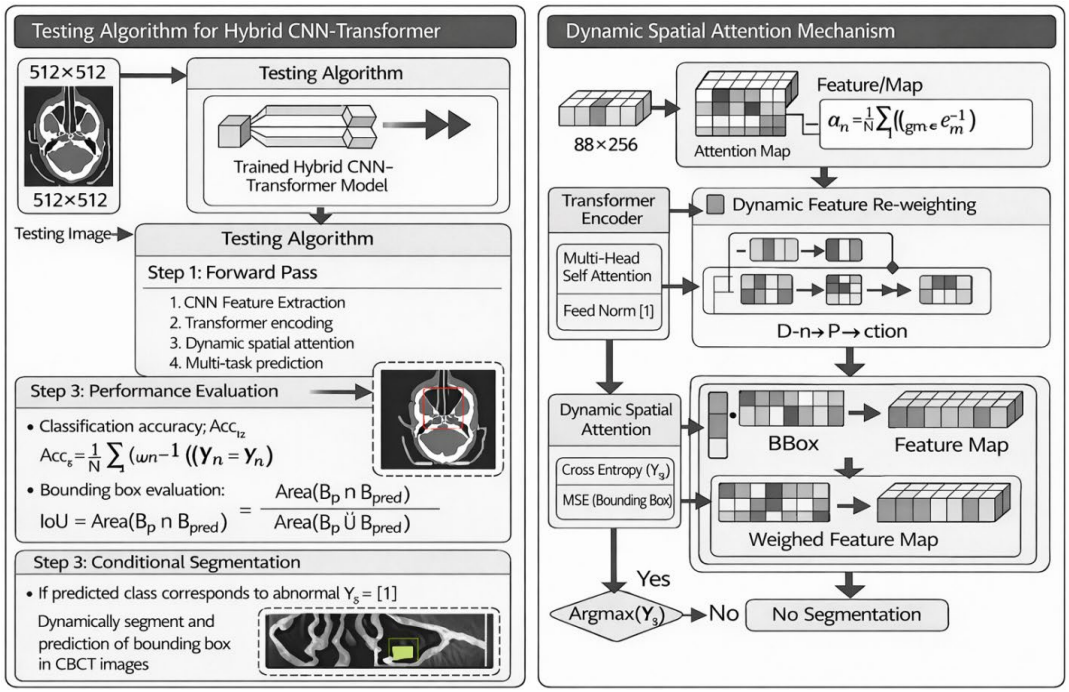


Figure 3. Testing architecture

3.3.1 CNN-based feature extraction in the hybrid model

The first stage of the 2D Hybrid CNN-Transformer model is focused on extracting spatial features from dental radiographs utilizing a CNN. This stage includes two convolutional blocks. Each block has a 2D layer of convolutional subsequently accompanied by applying batch normalization followed by a ReLU activation non-linearity. Convolutional layers operate with 3 × 3 kernels and a stride set to 1, the network capture fine details in the images. Batch normalization improve training stability and accelerate convergence. Let denote the input image X belongs to the space $R^{H\,XWXC}$, where W, C and H denote width, channels, and height, in the respective sequence. The CNN processes the input as follows:

$$F_1 = ReLU(\mathrm{BN}_1(\mathrm{Conv}_1(X))) \quad (1)$$

$$F_2 = ReLU(\mathrm{BN}_2(\mathrm{Conv}_2(F_1))) \quad (2)$$

Here, F_1 and F_2 are the intermediate and final feature maps. Each convolutional layer Conv_i has trainable weights and BN_i refers batch normalization layers.

Training Algorithm 1: Multi-task identification of abnormal positioned tooth in maxillary sinus-

Input

The training dataset comprises left and right maxillary sinus images, both containing and not containing abnormally positioned teeth.

$\mathcal{D}_{train} = \{(x_n, y^n_{cls}, y^n_{bbox})\}^{N_{train}}_{n=1}$

where

$x_n \in \mathbb{R}^{H\times W\times C}$- n-th CBCT image

$y^n_{cls} \in \{[1,0,0], [0,1,0], [0,0,1]\}$ - one-hot class label

$y^n_{bbox} = \{x_{min}, y_{min}, x_{max}, y_{max}\}$- Bounding box label for the n^{th}image Specifies the location of the abnormal tooth (if present).

Output

Trained hybrid CNN–Transformer model JC

Evaluation Metrics:

Classification Accuracy (Acc)

Bounding box IoU / MSE

Training Procedure

Step 1: Initialize Model Parameters

Initialize all trainable parameters:

CNN weights W_{conv}

Transformer weights W_M, W_N, W_V, W_O

Attention weights W_{att}

Classification head W_c

Detection head W_d

Set hyperparameters: $\eta, B, E, \lambda_1, \lambda_2$

Step 2: For each epoch $e = 1$to E

For each mini-batch $\{x_n\}^B_{n=1}$:

Step 3: Forward Propagation

(a) CNN Feature Extraction

The input image is passed through the CNN feature extractor as defined in Eqs. (1)-(2) and producing the final convolutional feature map F_2.

(b) Transformer Encoding

The extracted features are reshaped and processed by the Transformer encoder using the intra-attention mechanism described in Eqs. (3)-(8). The resulting context-enhanced representation is denoted as the output U′.

(c) Dynamic Spatial Attention (DSA)

The Dynamic Spatial Attention module refines the spatial representation according to Eqs. (9)-(10) yielding the spatially weighted feature representation Z′ .

(d) Multi-Task Prediction Heads

The refined features are fed into the classification and detection heads to obtain predictions using Eqs. (11)-(12). The outputs of the forward pass are therefore $\{\hat{y}_{cls}, \hat{y}_{bbox}\}$.

Step 4: Loss Computation

The multi-task loss function is computed as defined in Eqs. (13)-(15). The classification and bounding box regression losses are combined to obtain the total loss L.

Step 5: Backpropagation and Parameter Update

Model parameters are updated using the Adam optimization strategy described in Eq. (17) resulting in the updated parameter set $\Theta^{(t+1)}$

Repeat until convergence.

Testing algorithm2

Input:

Test dataset $\mathcal{D}_{test} = \{x_i\}^{N_{test}}_{i=1}$

Trained model JC

Output:

Predicted class labels and bounding boxes co-ordinates for the ithtest image Specifies the location of the abnormal

tooth (if present)= $\{\hat{y}^i_{cls}, \hat{y}^i_{bbox}\}^{N_{test}}_{i=1}$

Step 1: Forward Pass

For each test image x_i:

1.CNN feature extraction

2.Transformer encoding

3.Dynamic Spatial Attention

4.Multi-task prediction- Obtain = $\{\hat{y}^i_{cls}, \hat{y}^i_{bbox}\}$

Step 2: Performance Evaluation

Classification accuracy:

$Acc = \frac{1}{N_{test}} \sum^{N_{test}}_{i=1} \mathbb{1}(\arg\max\ \hat{y}^i_{cls} = y^i_{cls})$

Bounding box evaluation: $IoU = \frac{Area(B_{gt} \cap B_{pred})}{Area(B_{gt} \cup B_{pred})}$

Step 3: Conditional Segmentation

If predicted class corresponds to abnormal tooth= $\arg\max(\hat{y}_{cls}) \in$ {Left,Right}

then the predicted bounding box is used to extract the region of interest for fine-level segmentation.

3.3.2 Transformer-based learning of spatial dependencies

To overcome the limitation of capturing long-range dependencies within an image the output feature map from the CNNs denoted as $F \in R^{H' X W' \times D}$ is transformed into a sequential format and passed to a Transformer block [20]. The reshaping process converts the 2D feature map into a 1D sequence:

$$U = \text{Reshape}(F) \in R^{NXD} \text{where} N = H' \times W' \quad (3)$$

Each spatial location is treated as an individual token, allowing the intra-attention mechanism in the Transformer to model comprehensive spatial interactions in analyzing and representing complex anatomical structures across all parts of the image.

The Intra-attention mechanism is formulated as:

$$Intra - Attention(M, N, V) = softmax\left(\frac{MN^T}{\sqrt{d_n}}\right)V \quad (4)$$

The input sequence U is linearly transformed to generate the value V, query M and key N matrices using trainable weight parameters. These transformations are defined as:

$$V = UW_V, N = UW_N, M = UW_M \quad (5)$$

where, W_M, W_N, W_V are learnable projection matrices with dimensions R^{dXd_n}. The factor $\sqrt{d_n}$ scales the dot products for numerical stability of the softmax computation and n numbers of input tokens.

3.3.3 Multi-Head Intra-Attention

The Transformer architecture incorporates Multi-Head Intra-Attention (MHIA) to recognize spatial features across different regions. This approach processes the input across

several attention heads simultaneously, with each head attending to different aspects or regions of the data. The head_i is expressed as $\text{attention}(\text{UW}_M^{(i)}, \text{UW}_N^{(i)}, \text{UW}_V^{(i)})$ and the MHIA is formed by aggregating the outputs of all attention heads through concatenation, followed by projection with the learnable matrix W_0. In this setup, h represents the complete count of attention heads and $W_O \in \text{R}^{\text{hd}_k \times D}$ is a learnable matrix used to merge the outputs of each head. Every Transformer includes Feed-Forward Network (FFN) adds depth and non-linearity can be represented as:

$$FFN = W_2(Relu(\text{UW}_1 + b_1)) + b_2 \quad (6)$$

Dropout for regularization and Residual connections to support better gradient flow can be shown as:

$$U_{out} = LayerNorm(U + MHIA(U)) \quad (7)$$

$$\text{U}_{\text{final}} = \text{LayerNorm}(\text{U}_{\text{out}} + \text{FFN}(\text{U}_{\text{out}})) \quad (8)$$

The Transformer enhances the spatial feature representation learned by the CNN to detect complex abnormalities in medical images.

3.3.4 Dynamic Spatial Attention

The proposed framework introduces an enhanced hybrid architecture that combines convolutional neural networks and Transformer modules with a DSA mechanism for precise localization of abnormally positioned teeth within the maxillary sinus. Unlike traditional attention methods such as CBAM and non-local attention, which generate fixed or globally computed attention maps, the proposed DSA dynamically recalibrates spatial feature responses through context-driven weighting functions. This adaptive mechanism enables selective amplification of anatomically relevant regions while suppressing irrelevant background features is shown in Table 2.

Formally, the DSA module operates on intermediate feature maps by generating spatial attention weights conditioned on simultaneously capturing local features and global context, thereby improving discriminative capability. The framework is engineered to ensure complementary interaction between CNN-based local feature extraction, transformer-based global dependency modeling, and DSA-based spatial refinement. This integrated design significantly enhances localization accuracy and anatomical consistency, establishing a clear distinction from conventional hybrid models.

Table 2. Comparison of attention mechanisms

Feature	CBAM	Non-Local Attention	Proposed DSA
Attention Type	Channel + Spatial	Global Self-Attention	DSA
Adaptivity	Static (fixed structure)	Global dependency-based	Context-aware and adaptive
Focus	General feature refinement	Long-range relationships	Anatomical region-focused refinement
Computational Cost	Low	High	Moderate (optimized)
Localization Ability	Moderate	Good	High (task-specific)
Suitability for Medical Imaging	Limited	Moderate	High

Note: DSA = Dynamic Spatial Attention

Table 3. Layer-wise architecture and output specifications of the proposed CNN–transformer framework with Dynamic Spatial Attention

Stage	Operation	Output Size	Parameters
Input	Image	128 × 128 × 3	–
Conv block 1	Conv + Pool	64 × 64 × 32	896
Conv block 2	Conv + Pool	32 × 32 × 64	18496
Feature aggregation	Global Avg Pool	64	–
Transformer	Attention Block	128	8320
Classification head	Dense	1	129
Detection head	Dense	4	516
Total params	28357 (110.77 KB)		
Trainable params	28357 (110.77 KB)		
Non-trainable params	0 (0.00 Byte)		

In dental radiography different anatomical structures such as tooth exhibit varied spatial orientations and diagnostic relevance. To optimize feature selection a DSA module is integrated into the Hybrid CNN-Transformer architecture. The DSA module improves feature learning by focusing on the most significant spatial areas inside the extracted feature map. Initially, the feature map is reorganized into spatial representations, after which a learnable transformation is used to generate attention scores. These scores are normalized to identify and prioritize important areas, and the resulting weights are utilized on refine the original features. The updated features are then converted back to their spatial structure. Furthermore, incorporating a multi-head approach enables the model to capture multiple spatial dependencies, leading to enhanced performance. Assume the feature map extracted from the CNN-Transformer is denoted by $Z \in \text{R}^{\text{H}\times\text{W}\times\text{D}}$ in which H,W represents spatial dimensions and D represents the number of channels. Feature map information is refined by an attention mechanism that assigns spatial weights, highlighting the most relevant areas. This is achieved by applying a trainable projection to the flattened spatial feature map.

$$\text{A} = \text{softmax}(\text{Z}_{\text{flat}}.\text{W}_{\text{att}}) \quad (9)$$

where, W_{att} is a learnable parameter vector and a one-dimensional attention map that assigns importance to each spatial location, Z_{flat} is the reshaped feature map of size

$(H.W)X\,D$ .The softmax function normalize sum of the weights to 1. These learned attention scores are applied to modulate the original spatial features. This is performed via element-wise multiplication which enhances relevant regions and suppresses less important ones:

$$\acute{Z} = A \odot Z_{flat} \tag{10}$$

where, $\acute{Z}$ represents the refined feature map that can be reshaped back to its original dimensions enhances discriminative regions and aids in accurate classification and localization. By integrating these advanced deep learning methodologies into a unified end-to-end framework the proposed approach enhances detection accuracy contributing to the evolution of automated diagnostic tools in dental and maxillofacial radiology. Table 3 provides a detailed overview of the hybrid CNN and Transformer model architecture integrated with DSA outlining each layer's output dimensions, number of parameters and interconnections. It illustrates the progression from convolutional feature extraction to the Transformer module, culminating in the final classification and detection results.

3.4 Multi-task learning head framework

To further improve representational capacity, a multi-head attention strategy is adopted. Specifically, the channel dimension is partitioned into h subspaces, each of size $D_h = D/h$. For each head k, an independent attention map is computed over $Z_{flat}^{(k)} \in R^{(H.W)\times D_h}$ followed by feature reweighting. The predictions derived from all heads are concatenated to form the final refined feature representation. In this work, the attention head configuration is empirically set to h = 2, with standard Xavier initialization applied to all learnable parameters.

The proposed model is built as part of a multi-task learning paradigm facilitating concurrent optimization of classification and object detection tasks. The architecture integrates two specialized prediction heads each catering to a distinct objective. The classification head is designed to determine the presence of an abnormally positioned tooth within the input image. A terminal dense layer that applies softmax activation maps the derived features to a probability distribution across three target categories are normal, left maxillary sinus abnormality, and right maxillary sinus abnormality. Mathematically the classification output is expressed as:

$$\hat{y}_{cls} = softmax(W_c \acute{u}[0]) \tag{11}$$

The weight matrix $W_c \in R^{3\times D}$ of the classification head is trainable, $\acute{u}[0] \in R^D$ represents the first token from the refined Transformer output sequence u′, $\hat{y}_{cls} \in R^3$ yields the predicted probability distribution over the over the three classes are normal, left-sided abnormality, and right-sided abnormality. Conversely, the detection head is responsible for identifying the exact location of the abnormally positioned tooth by predicting its bounding box coordinates. The transformation of extracted features into normalized spatial coordinates is facilitated via a sigmoid-activated dense layer. The bounding box prediction is formulated as:

$$\hat{y}_{det} = sigmoid(w_d mean(\acute{u}, axis = 1)) \tag{12}$$

where, $mean(\acute{u}, axis = 1) \in R^D$ represents the average of all spatial tokens in the refined Transformer output, $W_d \in R^{4\,X\,D}$ is the trainable weight matrix for detection and $\hat{y}_{det} \in R^4$ corresponds to the normalized coordinates of the predicted bounding box is $(X_{min}, Y_{min}, X_{max}, Y_{max})$.

3.5 Multi-task loss optimization

The proposed hybrid CNN–Transformer framework is trained using a ramework for multi-task learning strategy designed to optimize concurrently both classification and localization objectives. The overall learning process is governed by a composite loss function that integrates loss terms for classification and bounding box prediction with task-specific weighting factors [21].

In the classification task, categorical CE loss is adopted to quantify the discrepancy relative to the predicted probability distribution together with the corresponding ground-truth one-hot encoded labels. This loss function is formulated as:

$$l_{cls} = -\sum_{c=1}^{3} y_c \log(\hat{y}_c) \tag{13}$$

where, y_c denotes the label of true class and $\hat{y}_c$ represents class c predicted probability obtained through the softmax activation layer.

For the localization task, the model minimizes the regression error between predicted versus actual bounding box locations using a MSE loss. The bounding box loss is formulated as:

$$l_{bbox} = \frac{1}{4}\sum_{j=1}^{4}(y_j^{bbox} - \hat{y}_j^{bbox})^2 \tag{14}$$

where, y_j^{bbox} and $\hat{y}_j^{bbox}$ correspond predicted against actual bounding box locations respectively.

The overall objective function used to train the model is calculated as a weighted linear aggregation of the two losses:

$$l = \lambda_1 l_{cls} + \lambda_2 l_{bbox} \tag{15}$$

where, λ_1 and λ_2 are positive scalar coefficients that control the relative contribution of classification and detection tasks during optimization. This balanced formulation enables the network to learn discriminative features for accurate abnormality classification while simultaneously improving localization precision.

3.6 Training configuration and optimization strategy

To ensure stable convergence and improved generalization, a set of training hyperparameters is carefully selected (Figure 4). The hyperparameter configuration is defined as:

$$H = \eta = 0.001, B = 32, E = 32, P_{drop} = 0.1 \tag{16}$$

where, η denotes the learning rate, B denotes the batch size, E denotes the total number of training epochs, and P_{drop} denotes the dropout probability.

Optimization is performed using the Adam algorithm, which adaptively updates each trainable parameter θ_i based on first- and second-order gradient moments. The parameter

update rule is expressed as:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \frac{\hat{m}_i}{\sqrt{\hat{v}_i} + \epsilon} \tag{17}$$

representing $\hat{m}_i$ and $\hat{v}_i$ represent the bias-corrected gradient first- and second-moment calculations, respectively, and ϵ is a small constant introduced for numerical stability.

To mitigate overfitting, dropout regularization is applied to intermediate feature activations during training:

$$\tilde{z}_i = z_i \cdot r_i, r_i \sim \text{Bernoulli}(1 - P_{drop}) \tag{18}$$

This mechanism randomly deactivates neurons with probability P_{drop} , thereby reducing co-adaptation among feature representations. Furthermore, Batch Normalization is employed following each convolutional or fully connected layer to stabilize learning dynamics:

$$BN(X) = \gamma \frac{X - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{19}$$

where, μ and σ^2 denote the mean and variance computed over the mini-batch, and γ and β are learnable parameters that scale and shift the activations. This normalization accelerates convergence and improves optimization stability.

Hyperparameters were determined empirically based on validation results. The framework is implemented in Built with TensorFlow and run on a workstation with an NVIDIA GPU, Intel i7 CPU, and 16 GB RAM. The model achieves an average inference time of approximately 0.05–0.1 seconds per CBCT image, indicating its suitability for real-world applications.

Collectively, the defined hyperparameter configuration and optimization strategy facilitate efficient learning and robust abnormality detection in dental radiographic images.

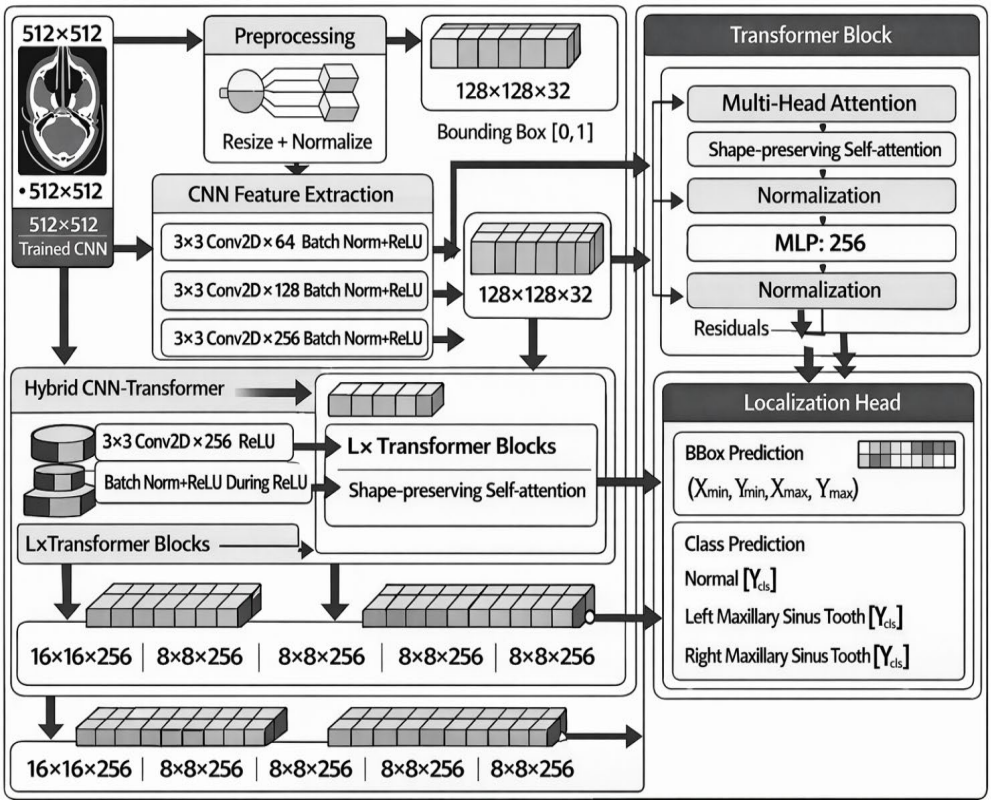


Figure 4. Layer details of the proposed method

3.7 Enhanced post-processing strategies for improved segmentation performance

To improve detection performance and maintain structural integrity in medical imaging tasks, we employ three synergistic post-processing techniques. Ensemble class probabilities combine outputs from multiple models to boost classification stability. Bounding box voting enhances localization accuracy by averaging overlapping predictions. Additionally, Adaptive soft non-maximum suppression (NMS) with anatomical awareness optimizes bounding box predictions by dynamically adjusting suppression based on anatomical context. These approaches significantly boost classification and detection precision, preserve topological consistency and improve clinical dependability. The ensemble approach minimizes model variance by integrating predictions from multiple models, resulting in more reliable and generalized performance. The voting scheme strengthens decision stability by combining outputs and choosing the most consistent result. Furthermore, NMS is utilized to discard overlapping and redundant detections, ensuring that only the most significant bounding boxes are retained, thereby enhancing detection precision.

3.7.1 Ensemble class probabilities

However, the proposed framework's accuracy can be impacted by variations in the input images such as changes in lighting, noise or the positioning of the tooth. These differences may lead to misclassifications when the tooth appears in a different orientation or brightness level compared to the training data. To improve robustness against such variations, an ensemble of predictions based on transformed

versions of the input image is used.

Let Q be the original input image, $AT = \{AT_1, AT_2 \ldots . AT_N\}$ be a set of N different augmentation transformations, the class probability for the ensemble model can be expressed as:

$$p_i = f\left(AT_i(Q)\right) \in R^C \tag{20}$$

where, f is the CNN and C is the number of classes.

The final predicted class ŷ is established by:

$$\hat{y} = \underbrace{\operatorname{argmax}}_{c} \widehat{p_c} \tag{21}$$

where, $\widehat{p_c}$ is the averaged probability for class c across all augmented versions.

3.7.2 Bounding box voting

Bounding box voting is a post-processing method in object detection used to refine multiple overlapping bounding boxes. When several detections exist for the same object, this technique combines them into a single, more accurate bounding box by calculating a weighted average based on their confidence scores. This helps reduce redundancy and enhances localization accuracy while preserving useful information from all predictions.

Let $b = \{bb_1, bb_2 \ldots . bb_n\}$ denote predicted bounding boxes set, where bb_i is an element of R^4 and is specified by the tuple $\{xb_{min}, xb_{max}, yb_{min}, yb_{max}\}$ and $S = \{s_1, s_2 \ldots . s_n\}$ be their corresponding confidence scores.

The final bounding box, denoted as:

$$bb_{vote} = \frac{\sum_{i=1}^{n} s_i . bb_i}{\sum_{i=1}^{n} s_i} \tag{22}$$

This approach results in a single, refined bounding box that better represents the object's location by leveraging the strengths of all overlapping predictions.

3.7.3 Anatomically-aware adaptive soft non-maximum suppression

For bounding box refinement, a soft NMS strategy is designed to incorporate anatomical priors. For each candidate box i, its confidence score s_i is adaptively updated using:

$$\acute{s}_i = s_i . \exp\left(-\gamma \cdot IoU_{ij}\right) . 1_{anat}(i) \tag{23}$$

Here $\acute{s}_i, s_i$ represents confidence score of original and refined detection i, γ- controlling suppression strength of decay factor, $1_{anat}(i)$ enforces anatomical consistency function is 1 if detection i satisfied otherwise 0 and IoU_{ij} is the intersection overunion with neighboring box j.

3.8 Abnormally positioned tooth in maxillary sinus segmentation

A classical image processing pipeline was designed for segmentation and boundary delineation structures of tooth in grayscale maxillofacial radiographic images. The image was initially reshaped to a standardized spatial resolution of 256 × 256 pixels to ensure consistency across processing steps. To prepare for segmentation grayscale normalization was performed followed by global thresholding to isolate hyperdense regions due to tooth structures high radio density in CT imaging. Morphological operations specifically binary opening using a 3 × 3 kernel were applied to suppress small-scale noise and enhance the continuity of anatomical boundaries [22]. Contour extraction was subsequently performed using connected component analysis and the largest connected contour assumed to represent the primary tooth structure was selected [23]. This contour was overlaid onto the original image in green to facilitate visual inspection of the segmented boundary.

Let $I(x, y) \in R^{H X W}$ where $I(x, y)$ represent the original grayscale dental radiograph of height H and width. To ensure uniformity in subsequent processing, the image is resized to a fixed resolution:

$$I_r\left(x, y\right) = Resize\left(I\left(x, y\right), 256 \times 256\right) \tag{24}$$

where, I_r denotes the resampled image. The intensity levels transformed to values between 0 and 1 to facilitate threshold-based operations:

$$I_n\left(x, y\right) = \frac{I_r\left(x, y\right)}{255} \tag{25}$$

A fixed global threshold $T \in [0,1]$ is applied to derive a binary mask, isolating regions of high intensity, which are likely to correspond to tooth structures:

$$B(x, y) = \begin{cases} 1, \text{if } I_n(x, y) \geq T \\ 0, \text{otherwise} \end{cases} \tag{26}$$

To enhance the segmentation by eliminating small noise artifacts, a morphological opening operation is performed:

$$B_{clean}\left(x, y\right) = \left(B \; o \; K\right)(x, y) \tag{27}$$

where, o denotes the morphological opening and $K \in \{0,1\}\{0,1\}^{3 \times 3}$ is a binary structuring element of size 3 × 3. Contour extraction is conducted on the denoised binary mask B_{clean} yielding a set of candidate contours $C = \{C_1, C_2, \ldots C_N\}$. Among these, the contour with the maximum area is selected as the region of interest:

$$C_{max} = \arg \underbrace{\max}_{C_i \in C} Area(C_i) \tag{28}$$

The final segmentation output is generated by overlaying the selected contour C_{max} onto the original image.

$$I_{seg} = DrawContour(I_r, C_{max}, color = green) \tag{29}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Ablation analysis

To ensure reliable performance despite a limited dataset, 5-fold cross-validation was applied on 300 stratified CBCT scans, along with data augmentation including rotation, scaling, and intensity variations to enhance dataset diversity and mitigate overfitting. An ablation study was conducted to assess the contribution of each module within the hybrid CNN–Transformer framework shown in Table 4. Results indicate that the CNN backbone provides robust local feature

extraction, the Transformer module enhances global contextual modeling, and the DSA module improves spatial feature refinement. The full model achieved the best performance with 94% classification accuracy, 87.5% precision, 93.3% recall, and 88.7% mAP, demonstrating that the observed improvements result from the synergistic combination of all components rather than any single module.

Table 4. Ablation study of model components and their impact on performance

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	mAP (%)
CNN only	83.2	82.5	80	80.1
CNN + Transformer	85	85	82.2	83.5
CNN + DSA	86.5	84	84.5	82.8
CNN + Transformer + DSA	88	87.5	86	88.7

4.2 Quantitative impact of integrated post-processing techniques

Table 5. Quantitative comparison of identification of abnormally positioned tooth in maxillary sinus performance with post-processing techniques

No.	Method	IoU	MSE ↓	Precision	Recall
1	Propsed 2D Hybrid CNN–Transformer with DSA	78	0.021	87.5	86
2	+ Ensemble Class Probabilities	82	0.018	89.5	88
3	+ Bounding Box Voting	84	0.016	91.5	90
4	+ Anatomically-aware Soft NMS	86	0.013	93.5	92

Note: MSE = Mean Squared Error

The incorporation of integrated post-processing strategies significantly improves detection robustness and spatial localization accuracy. As summarized in Table 5, the baseline CNN–Transformer with DSA achieves an IoU of 0.78 for bounding box localization. The addition of ensemble probability aggregation increases IoU to 0.83. Further refinement through bounding box voting and adaptive soft NMS progressively enhances localization performance, achieving a final IoU of 0.85. These observations imply that the proposed post-processing framework effectively reduces redundant detections and improves spatial alignment of predicted bounding boxes, thereby enhancing anatomical consistency in identifying abnormally positioned tooth within the maxillary sinus.

4.3 Proposed method-identification of abnormally positioned tooth in maxillary sinus and segmentation results

The full operational workflow of the proposed framework is illustrated through the training and testing phases. During the training phase, the network learns to perform joint classification and bounding box regression, as shown in Figure 5.

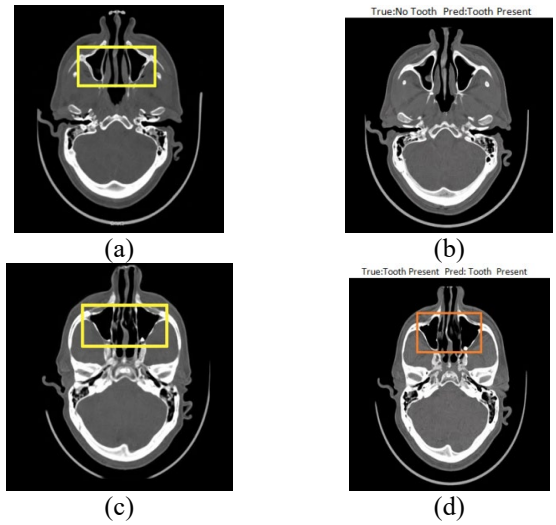


Figure 5. Visualization of training data, (a) and (c) Ground-truth annotations CBCT images (b) and (d) Corresponding predictions

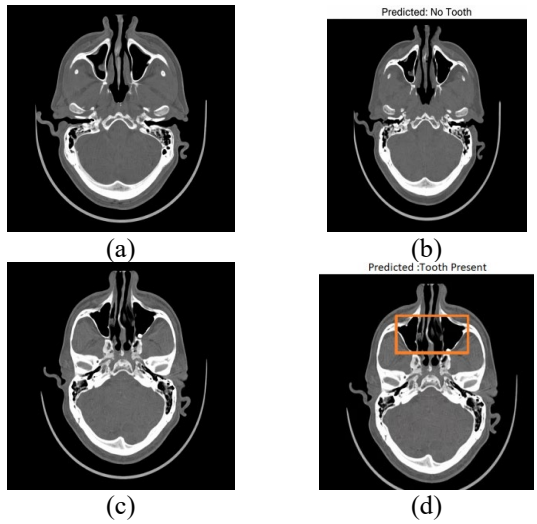
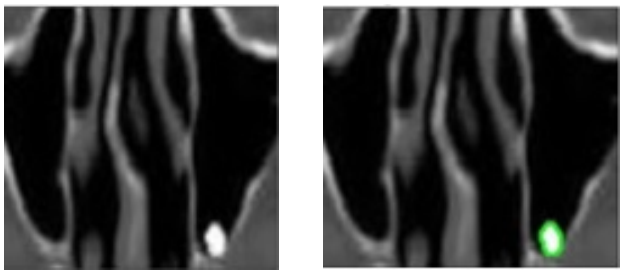


Figure 6. Classification and detection (a) and (c) Testing CBCT images (b) and (d) corresponding predicted image

Visualization of training data, Figures 5(a) and (c) present resized CBCT input images without and with an abnormally positioned tooth, respectively, along with ground-truth bounding box annotations. Figures 5(b) and (d) show the corresponding model predictions after training, where the predicted bounding boxes indicate the localized position of the abnormal tooth positioned inside the maxillary sinus. The testing phase is demonstrated in Figure 6. Figures 6(a) and (c) show representative unseen CBCT images, including a case with an abnormally positioned tooth in the left maxillary sinus. The corresponding predictions are illustrated in Figures 6(b) and (d), where the model simultaneously performs classification (tooth present/absent) and detection, highlighting the localized region using a predicted bounding box.

Following detection, a region of interest is extracted based on the predicted bounding box shown in Figure 7. Detection and segmentation, Figure 7(a) isolates the predicted area for focused analysis. Segmentation is then performed within this

ROI to refine structural delineation. The resulting segmentation map Figure 6(b) accurately outlines the abnormal tooth at the pixel level, enabling precise anatomical boundary identification through contour visualization. This end-to-end multi-task framework integrates convolutional feature extraction followed by Transformer-based modeling of global context, and DSA to ensure robust performance under complex anatomical variations. By combining classification, detection, and segmentation within a unified architecture, the proposed method provides clinically meaningful outputs that can support diagnostic decision-making and surgical planning in dental and maxillofacial radiology.



(a) (b)

Figure 7. Detection and segmentation results (a) ROI selection highlighting the region of interest; (b) segmentation of an abnormally positioned tooth in the left maxillary sinus

4.4 Evaluation and performance metrics

This section outlines the evaluation framework and performance assessment pertaining to the proposed deep learning model for detecting abnormally positioned tooth in the maxillary sinus. The analysis encompasses both training and testing stages to examine optimization stability and model generalization. Throughout the training process, learning behavior was monitored by observing loss convergence patterns and progressive enhancement of evaluation metrics. The steady decline in total loss reflects effective optimization and stable parameter updates. For performance validation, the trained model was tested on independent CBCT datasets to measure its generalization capability. The framework demonstrated a classification accuracy of 94%, highlighting the model's strength in discriminating between normal and abnormal tooth positioning within the maxillary sinus. Regarding detection performance, the model attaining a precision of 87.5%, demonstrating a recall of 93.3%, demonstrating reliable localization of abnormal tooth while maintaining a controlled false-positive rate. Additionally, the object detection evaluation produced a mean Average Precision (mAP) of 88.7%. A comparative study with baseline architectures, including ResNet50 integrated with YOLOv5 and U-Net combined with CNN models, is summarized in Table 6.

The proposed Hybrid CNN–Transformer with DSA achieves improved overall results in both classification accuracy and detection mAP. Figure 8 presents a graphical comparison of classification and detection metrics between the proposed method and baseline models, further illustrating the balanced and effective performance of the multi-task framework.

To investigate the efficacy of statistical relevance of the observed performance improvements, a paired t-test was performed between the proposed model and baseline methods using Dice scores from the test set shown in Table 7. The null hypothesis assumes no meaningful difference between the models. The findings suggest that the proposed approach attains statistically significant improvements ($p < 0.05$), confirming that the gains are not due to random variation but represent true performance enhancement.

Table 6. Comparative analysis of proposed method and baseline models

Model	Classification Accuracy	Object Detection (mAP)
ResNet50 with YOLOv5	83.3%	80.5%
U-Net with CNN	87%	82.2%
TransUNet (Transformer-based)	89%	84%
Swin-Unet (Transformer-based)	91%	85%
Proposed 2D Hybrid CNN–Transformer with DSA	**94%**	**88.7%**

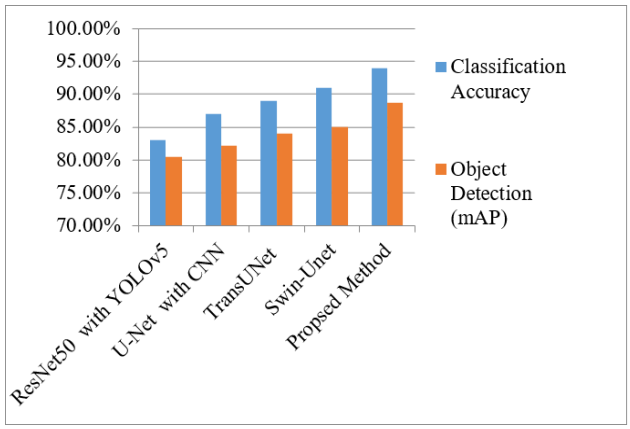


Figure 8. Graphical comparison of classification accuracy and object detection performance with baseline models

Table 7. Statistical significance analysis (Paired t-test)

Comparison	Accuracy (%)	mAP (%)	P-Value	Significance
Proposed vs ResNet50 + YOLOv5	94 vs 83.3	88.7 vs. 80.5	0.008	Significant ($p < 0.05$)
Proposed vs U-Net (CNN)	94 vs 87	88.7 vs. 82.2	0.015	Significant ($p < 0.05$)
Proposed vs TransUNet	94 vs 89–91	88.7 vs. 84–86	0.021	Significant ($p < 0.05$)
Proposed vs Swin-Unet	94 vs 90–92	88.7 vs. 85–87	0.028	Significant ($p < 0.05$)

5. CONCLUSION

This research work presented a deep learning–guided a framework designed for automated identification of abnormally positioned tooth in the maxillary sinus using a Hybrid CNN–Transformer architecture integrated with DSA. The proposed model achieved high classification accuracy and strong detection performance, demonstrating its capability to reliably differentiate normal from abnormal cases while

accurately localizing pathological tooth. The hybrid architecture effectively captured both local spatial features and long-range contextual dependencies, enabling improved representation of complex anatomical structures in CBCT images. The inclusion of DSA enhanced feature refinement by emphasizing clinically relevant regions, thereby reducing false positives and improving interpretability. Experimental results confirmed that the proposed approach outperformed conventional CNN-based models in both classification and detection tasks. Furthermore, the model exhibited stable generalization across variations in image quality, anatomical complexity, and patient-specific differences. Its computational efficiency and rapid inference time highlight its feasibility for integration into real-time computer-aided diagnostic systems. Overall, the combination of CBCT imaging and AI-driven multi-task learning offers a reliable and efficient tool to support diagnosis, surgical planning, and clinical decision-making in cases involving abnormally positioned tooth within the maxillary sinus.

6. FUTURE WORK

This study is limited by the relatively small dataset size, primarily due to the rarity of such cases and the requirement for expert annotation of CBCT images. Future studies will aim to further improve framework guided by deep learning for automated proposed Hybrid CNN–Transformer framework across heterogeneous patient populations, imaging systems, and acquisition settings. Increasing the dataset size to include a wider range of anatomical variations and pathological presentations will contribute to enhancing model stability and reliability. Further evaluation is necessary to assess model performance in complex clinical scenarios, particularly when abnormal tooth is located adjacent to critical anatomical structures or accompanied by secondary pathological conditions.

In addition, multi-center validation studies should be conducted to verify reproducibility and ensure consistent performance across independent CBCT datasets. From a technical standpoint, extending the current frame work to three-dimensional CBCT analysis using 3D CNN–Transformer architectures may enable more comprehensive volumetric feature learning. The incorporation of anatomically guided constraints and uncertainty estimation mechanisms could also improve interpretability and strengthen clinical confidence in automated predictions.

Moreover, future work will explore the deployment of the proposed model within an integrated clinical decision-support system to facilitate seamless incorporation into radiology workflows. Long-term clinical investigations will be essential to determine the practical impact of AI-assisted diagnosis on treatment planning efficiency, surgical outcomes, and overall patient care. Subsequent research will concentrate on augmenting the dataset through multi-center sources collaboration to improve generalization, evaluating the model on complex clinical cases, and extending the framework to 3D CBCT analysis for volumetric feature learning. Multi-center validation and uncertainty estimation will also be incorporated to enhance reproducibility, interpretability, and clinical reliability.

REFERENCE

[1] Irimia, Ó.A., Dorado, C.B., Marino, J.A.S., Rodríguez, N.M., González, J.M.M. (2010). Meta-analisis of the etiology of odontogenic maxillary sinusitis. Medicina Oral, Patología Oral Y Cirugía Bucal. Ed. Inglesa, 15(1): 16.

[2] Mehra, P., Murad, H. (2004). Maxillary sinus disease of odontogenic origin. Otolaryngologic Clinics of North America, 37(2): 347-364. https://doi.org/10.1016/S0030-6665(03)00171-3

[3] Yan, Y., Li, J., Zhu, H., Liu, J., Ren, J., Zou, L. (2021). CBCT evaluation of root canal morphology and anatomical relationship of root of maxillary second premolar to maxillary sinus in a western Chinese population. BMC Oral Health, 21(1): 358. https://doi.org/10.1186/s12903-021-01714-w

[4] Dinç, K., İçöz, D. (2024). Maxillary sinus volume changes in individuals with different craniofacial skeletal patterns: CBCT study. BMC Oral Health, 24(1): 1516. https://doi.org/10.1186/s12903-024-05341-z

[5] Themkumkwun, S., Kitisubkanchana, J., Waikakul, A., Boonsiriseth, K. (2019). Maxillary molar root protrusion into the maxillary sinus: A comparison of cone beam computed tomography and panoramic findings. International Journal of Oral and Maxillofacial Surgery, 48(12): 1570-1576. https://doi.org/10.1016/j.ijom.2019.06.011

[6] Ngoc, V.T.N., Duc, N.M., Dinh, T.C., Dinh, T.C. (2019). Cone beam computed tomography application in finding ectopic tooth: A systemic analysis and a case report. Open Access Macedonian Journal of Medical Sciences, 7(24): 4333. https://doi.org/10.3889/oamjms.2019.386

[7] Nashef, A., Joachim, M.V., Liubin, N., Raziq, M.A., El-Naaj, I.A., Laviv, A. (2025). The modified Caldwell-Luc approach for treating odontogenic maxillary sinusitis without need for functional endoscopic sinus surgery: A retrospective study. Journal of Oral and Maxillofacial Surgery, 83(2): 199-207. https://doi.org/10.1016/j.joms.2024.09.006

[8] Wu, Z., Yu, X., Chen, Y., Chen, X., Xu, C. (2024). Deep learning in the diagnosis of maxillary sinus diseases: A systematic review. Dentomaxillofacial Radiology, 53(6): 354-362. https://doi.org/10.1093/dmfr/twae031

[9] Sivari, E., Senirkentli, G.B., Bostanci, E., Guzel, M.S., Acici, K., Asuroglu, T. (2023). Deep learning in diagnosis of dental anomalies and diseases: A systematic review. Diagnostics, 13(15): 2512. https://doi.org/10.3390/diagnostics13152512

[10] Aktuna Belgin, C., Kurbanova, A., Aksoy, S., Akkaya, N., Orhan, K. (2025). Detection of maxillary sinus pathologies using deep learning algorithms. European Archives of Oto-Rhino-Laryngology, 282(9): 4727-4734. https://doi.org/10.1007/s00405-025-09451-4

[11] Fan, W., Zhang, J., Wang, N., Li, J., Hu, L. (2023). The application of deep learning on CBCT in dentistry. Diagnostics, 13(12): 2056. https://doi.org/10.3390/diagnostics13122056

[12] Liu, Y., Han, L., Yao, B., Li, Q. (2024). STA-Former: Enhancing medical image segmentation with Shrinkage Triplet Attention in a hybrid CNN-Transformer model. Signal, Image and Video Processing, 18(2): 1901-1910. https://doi.org/10.1007/s11760-023-02893-5

[13] Fan, Y., Song, J., Yuan, L., Jia, Y. (2025). HCT-Unet: Multi-target medical image segmentation via a hybrid CNN-transformer Unet incorporating multi-axis gated multi-layer perceptron. The Visual Computer, 41(5):

3457-3472. https://doi.org/10.1007/s00371-024-03612-y
[14] Shehzadi, T., Hashmi, K.A., Liwicki, M., Stricker, D., Afzal, M.Z. (2025). Object detection with transformers: A review. Sensors, 25(19): 6025. https://doi.org/10.3390/s25196025
[15] Chen, Y.L., Lin, C.L., Lin, Y.C., Chen, T.C. (2024). Transformer-CNN for small image object detection. Signal Processing: Image Communication, 129: 117194. https://doi.org/10.1016/j.image.2024.117194
[16] Arkin, E., Yadikar, N., Xu, X., Aysa, A., Ubul, K. (2023). A survey: Object detection methods from CNN to transformer. Multimedia Tools and Applications, 82(14): 21353-21383. https://doi.org/10.1007/s11042-022-13801-3
[17] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615
[18] Bayrakdar, I.S., Elfayome, N.S., Hussien, R.A., Gulsen, I.T., et al. (2024). Artificial intelligence system for automatic maxillary sinus segmentation on cone beam computed tomography images. Dentomaxillofacial Radiology, 53(4): 256-266. https://doi.org/10.1093/dmfr/twae012
[19] Altun, O., Özen, D.Ç., Duman, Ş.B., Dedeoğlu, N., et al. (2024). Automatic maxillary sinus segmentation and pathology classification on cone-beam computed tomographic images using deep learning. BMC Oral Health, 24(1): 1208. https://doi.org/10.1186/s12903-024-04924-0
[20] Chen, J. (2025). Convolutional neural network for maxillary sinus segmentation based on the U-Net architecture at different planes in the Chinese population: A semantic segmentation study. BMC Oral Health, 25(1): 961. https://doi.org/10.1186/s12903-025-06408-1
[21] Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L. (2022). Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized Medical Imaging and Graphics, 95: 102026. https://doi.org/10.1016/j.compmedimag.2021.102026
[22] Jung, S.K., Lim, H.K., Lee, S., Cho, Y., Song, I.S. (2021). Deep active learning for automatic segmentation of maxillary sinus lesions using a convolutional neural network. Diagnostics, 11(4): 688. https://doi.org/10.3390/diagnostics11040688
[23] Yoo, Y.S., Kim, D., Yang, S., Kang, S.R., et al. (2023). Comparison of 2D, 2.5D, and 3D segmentation networks for maxillary sinuses and lesions in CBCT images. BMC Oral Health, 23(1): 866. https://doi.org/10.1186/s12903-023-03607-6