computers

MDPI

Article

Hybrid CNN-Swin Transformer Model to Advance the Diagnosis of Maxillary Sinus Abnormalities on CT Images Using Explainable AI

Mohammad Alhumaid [1,2,*] and Ayman G. Fayoumi [1]

[1] Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; afayoumi@kau.edu.sa
[2] College of Computer Science and Engineering, University of Hail, Hail 81481, Saudi Arabia
* Correspondence: malhumaid0003@stu.kau.edu.sa or m.alhomaid@uoh.edu.sa

Abstract

Accurate diagnosis of sinusitis is essential due to its widespread prevalence and its considerable impact on patient quality of life. While multiple imaging techniques are available for detecting maxillary sinus, computed tomography (CT) remains the preferred modality because of its high sensitivity and spatial resolution. Although recent advances in deep learning have led to the development of automated methods for sinusitis classification, many existing models perform poorly in the presence of complex pathological features and offer limited interpretability, which hinders their integration into clinical workflows. In this study, we propose a hybrid deep learning framework that combines EfficientNetB0, a convolutional neural network, with the Swin Transformer, a vision transformer, to improve feature representation. An attention-based fusion module is used to integrate both local and global information, thereby enhancing diagnostic accuracy. To improve transparency and support clinical adoption, the model incorporates explainable artificial intelligence (XAI) techniques using Gradient-weighted Class Activation Mapping (Grad-CAM). This allows for visualization of the regions influencing the model's predictions, helping radiologists assess the clinical relevance of the results. We evaluate the proposed method on a curated maxillary sinus CT dataset covering four diagnostic categories: Normal, Opacified, Polyposis, and Retention Cysts. The model achieves a classification accuracy of 95.83%, with precision, recall, and *F1 score* all at 95%. Grad-CAM visualizations indicate that the model consistently focuses on clinically significant regions of the sinus anatomy, supporting its potential utility as a reliable diagnostic aid in medical practice.

Keywords: maxillary sinus; deep learning; vision transformer; grad-cam

check for updates

Academic Editor: Rafiqul Chowdhury

Received: 10 September 2025
Revised: 21 September 2025
Accepted: 29 September 2025
Published: 2 October 2025

Citation: Alhumaid, M.; Fayoumi, A.G. Hybrid CNN-Swin Transformer Model to Advance the Diagnosis of Maxillary Sinus Abnormalities on CT Images Using Explainable AI. *Computers* **2025**, *14*, 419. https://doi.org/10.3390/computers14100419

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Sinusitis refers to an inflammation or swelling of the internal tissues inside the paranasal cavity. It is a common medical condition that affects a significant population worldwide [1], contributing to the loss of productivity among individuals as well as leading to a substantial socio-economic burden due to healthcare consumption. According to an estimate by EUFOREA in 2018, a total of 10% of the population in Europe suffered from chronic rhinosinusitis (CRS) [2]. In a survey conducted in the USA, around 14.7% of the participants were reported to have suffered from sinusitis [3]. A recent study [4] performed with 3602 participants from different regions of Saudi Arabia showed that 26.3% of individuals (75.1% being female) were diagnosed with CRS. The common risk factors for

sinusitis include upper respiratory infections, nasal blockage, allergies, asthma, a deviated septum, and a weakened immune system.

The term maxillary sinusitis is associated with inflammation of the maxillary sinus inside the paranasal region. Located within the maxilla and adjacent to the nasal cavity, the maxillary sinus is the largest among the paranasal sinuses. It plays a pivotal role in maintaining sinus health and the understanding of sinus-related pathologies [5]. Maxillary sinusitis can be classified as acute or chronic [6] depending upon the clinical symptoms. Acute sinusitis lasts up to four weeks, with symptoms like nasal congestion, purulent discharge, and facial pain. Chronic sinusitis can be caused by viruses, bacteria, or fungi, and persists for more than 12 weeks, marked by prolonged inflammation, nasal polyps, and recurrent infections.

For an accurate diagnosis of the maxillary sinusitis, physicians often rely on different imaging modalities, which include conventional radiography (X-rays) [7], Magnetic Resonance Imaging (MRI) [8], Computed Tomography (CT) [9], ultrasound imaging [10], and endoscopy [11]. CT is the gold standard for diagnosing sinus diseases due to its high sensitivity and ability to detect soft and bone tissues, enabling early detection and prevention of serious maxillary sinusitis complications [12]. Polyposis, retention cysts, mucosal thickening and air fluid levels are the types of CT findings on maxillary sinus (MS) images [13]. However, the anatomical structure of the maxillary sinus area makes it challenging to distinguish these conditions. The similar appearance of retention cysts and opacified MS, or the minor mucosal thickening, can make it very difficult to accurately differentiate these conditions unless highly advanced image analysis is performed.

Diagnosis for maxillary sinusitis does not only involve the detection, but the intensity of the disease also needs to be assessed for proper treatment [5]. Normally, this is conducted manually by a skilled radiologist, which is time consuming, laborious and requires intensive expertise. For that purpose, traditional machine learning (ML) models have been deployed to automate and make the process efficient [14] to assist radiologists in diagnosing maxillary sinusitis with accuracy; unfortunately, the existing ML models suffer from the fact that we can only use direct output features and are incapable of extracting the implicit features present in raw imaging data [9,15,16]. Different modalities have also been proposed using the existing approaches, such as X-ray [17,18], MRI [8], ultrasound [10], and endoscopy [11]; however, CT imaging [8,9,19] is considered more preferable as it has better accuracy, sensitivity and suitability for complex cases.

The use of deep learning (DL) approaches has recently gained popularity as they have shown good performance in learning meaningful information and radiomic features [20]. In addition, image segmentation techniques possess high potential in enhancing diagnosis of sinusitis through aiding detection methods effectively [21,22]. Several deep learning-based approaches have been proposed for the diagnosis and analysis of sinusitis. A number of works have been reported demonstrating the potential of deep learning for diagnosis of maxillary sinusitis with radiographs and X-ray images [17,23,24]. Similarly, a number of method have been proposed for CT modality, which used standard convolutional neural networks (CNN) based models [25] or adaptations of CNNs, e.g., 3D-CNN [26], Aux-MVNeT [18] and SinusC-Net [27]. Recent years have witnessed an increasing interest in vision transformers (ViT) for various medical imaging applications, demonstrating highly encouraging performance [28,29]; however, the use of ViT is still largely unexplored in applications involving diagnosis and analysis of maxillary sinusitis.

Hence, this paper proposes a hybrid deep learning framework combining convolutional neural networks (CNN) and Swin Transformers (ViT) to precisely classify maxillary sinus abnormalities from CT images. Notably, few CNN-based models for sinus detection have been reported in the literature in recent years; however, the majority of these

works employ a single model for classification. Additionally, the lack of explainability in many deep learning models poses a significant barrier to their clinical adoption, as radiologists often require transparent reasoning for model predictions. Accordingly, the main contributions of this study are as follows:

i. Proposes a hybrid deep learning model combining CNN and Swin Transformer to improve the classification of maxillary sinus abnormalities from CT images.
ii. Utilizes the strengths of CNNs for capturing local features and Swin Transformers for modeling long-range dependencies in sinus imaging.
iii. Integrates an Explainable AI (XAI) technique, specifically Gradient-weighted Class Activation Mapping (Grad-CAM), to enhance transparency and interpretability of the model's decisions.

The structure of this paper is organized as follows: Section 2 presents the related work, highlighting existing approaches for maxillary sinus detection and classification using machine learning and deep learning techniques. Section 3 describes the proposed methodology, including data collection, preprocessing, model architecture, training procedures, and hyperparameter tuning. Section 4 details the experimental results and performance evaluation of the model. Section 5 discusses the application of Grad-CAM to provide visual interpretability of model predictions. Finally, Section 6 concludes the paper and outlines potential directions for future research.

2. Related Works

Several methods have been reported in the literature for automated detection of maxillary sinusitis, with the majority relying on convolutional neural networks (CNNs). Jeon et al. [23] presented a CNN-based approach for the detection and classification of frontal, ethmoid, and maxillary sinusitis using Waters' and Caldwell's radiographs. Laura et al. [19] proposed an ensemble approach that combined Darknet-19 deep neural network with YOLO for the detection of nasal sinuses and cavities with CT images. Kim et al. [17] also developed an ensemble method built on multiple CNN models (VGG-16, VGG-19, ResNet-101) for the detection of sinusitis using a majority voting approach with X-ray images, achieving encouraging results. Ozbay and Tunc [25] also made a useful contribution by proposing a method that relied on thresholding-based CT image segmentation Otsu's method [30] followed by classification of sinus abnormalities using a CNN-based model with promising performance; however, small data dataset led to generalization issues for diverse images.

Likewise, more sophisticated methods employed advanced architectures for diagnosing and screening of Maxillary sinusitis. Lim et al. [18] presented an auxiliary classifier-based multi-view CNN model, called Aux-MVNet, that was aimed at the localization of maxillary sinusitis and classification of severity levels using X-ray images. Çelebi et al. [28] developed a Swin Transformer-based architecture for maxillary sinus detection by utilizing the window multi-head self-attention mechanism in CBCT images. Kuwana et al. [24] focused on the detection and classification of sinus lesions into healthy and inflamed categories using DetectNet model with panoramic radiographs. Murata et al. [31] employed AlexNet for the diagnosis of maxillary sinusitis on panoramic radiographic image dataset created using varying data augmentation techniques. In addition, some studies explored the use of deep learning approaches for sinusitis screening using 3D volumetric datasets. Hwang et al. [27] introduced the SinusC-Net model, a 3D distance-guided network, for surgical plan classification for maxillary sinus augmentation on CBCT images. Likewise, Bhattacharya et al. [26] demonstrated robust classification performance using a 3D CNN model with supervised contrastive loss to classify sinus volumes into normal and abnormal

categories. These sophisticated methods resulted in improved performance across different imaging modalities; however, they failed to generalize in constrained environments.

Similarly, there exist approaches employing transfer learning, i.e., customized architecture built on top of pretrained models to perform sinusitis diagnosis in an effective and efficient manner. Mori et al. [32] proposed one such method for a robust detection and diagnosis of maxillary sinusitis on panoramic radiographs. Similarly, Kotaki et al. [33] also investigated the use of transfer learning to enhance the performance of the diagnosis of maxillary sinusitis with radiography. However, fine-tuning of pretrained model for maxillary diagnosis again requires a significant volume of image data. Moreover, Oğuzhan Altun et al. [34] developed a modified YOLOv5x architecture with transfer learning for automated segmentation of maxillary sinuses and associated pathologies in CBCT images, achieving high accuracy and precision, although the study's small dataset of 307 images limits generalizability. Similarly, Ibrahim Sevki Bayrakdar et al. [35] employed the nnU-Net v2 model for automatic segmentation of maxillary sinuses in CBCT volumes, showing strong performance with a limited dataset and manual annotations. While the use of nnU-Net v2 is a strength, the small dataset and lack of external validation are limitations for broader clinical applicability. An interesting work [36] demonstrated the use of Swin transformers for maxillary sinus detection, but it employed cone beam computed tomography (CBCT), which has different characteristics compared to traditional CT scans, particularly in terms of resolution and the types of images it produces. This limitation makes it less generalizable to standard CT-based sinusitis detection, and highlights the need for further exploration of transformer-based models with traditional CT data to assess their viability in real-world clinical settings. Additionally, the study lacked explainability mechanisms, limiting its clinical transparency, and it did not assess performance across different stages of the disease, which is important for reliable diagnostic support.

As reviewed above, significant advancement has been made in terms of automated medical image analysis [16], including sinusitis diagnosis and severity identification. Several deep learning-based architectures have been investigated to perform an automated diagnosis and classification of maxillary sinusitis [26,27,31]. However, there is a need to explore the potential of ViT architectures for maxillary classification as these are being extensively used for medical image analysis [28,29]. Table 1 presents a summary of related works, highlighting the need for hybrid architectures that leverage the strengths of both convolutional neural networks and transformer-based models to improve diagnostic accuracy, interpretability, and generalizability across diverse imaging modalities.

Table 1. Summary of the existing related approaches for maxillary sinusitis diagnosis and classification.

Reference	Type of AI Model	Method/Model	Explainable AI	Imaging Modality	Problem	Limitation
[23]	Deep Learning	CNN	No	Radiograph images	Classify	Limited to 2D views; lacks generalization
[19]	Transformer-based	Darknet-19 + YOLO	No	CT images	Detect & Classify	High complexity; no interpretability
[17]	Deep Learning	VGG-16, VGG-19 and ResNet101	No	X-ray images	Classify	Dependent on voting; slow inference

Table 1. *Cont.*

Reference	Type of AI Model	Method/Model	Explainable AI	Imaging Modality	Problem	Limitation
[25]	Deep Learning	CNN	No	CT images	Classify	Limited diagnostic insight; lacks interpretability
[18]	Multi-view CNN	AuX-MVNet	No	X-ray images	Localize & Classify	Requires multiple views; lacks explainability
[36]	Transformer-based	Swin Transformer	No	CBCT images	Detect	High complexity; no interpretability
[24]	Transformer-based	DetectNet	No	Panoramic radiograph images	Detect & Classify	Binary classification only; lacks granularity
[31]	Transformer-based	AlexNet	No	CBCT images	Classify	Shallow network; low feature depth
[27]	Transformer-based	SinusC_Net	No	MRI image	Classify	Architecturally complex; lacks clinical interpretability
[26]	Deep Learning	3D-CNN	No	MRI image	Classify	High computation; lacks interpretability
[32]	Deep Learning	DNN	No	Panoramic radiograph images	Detect & Classify	limited transparency for clinical use
[33]	Deep Learning	DNN	No	Radiograph images	Detect & Classify	Limited fine-tuning capability
[34]	Deep Learning	YOLOv5x with transfer learning	No	CBCT images	Detect & Classify	Lacks interpretability in complex cases
[35]	Deep Learning	nnU-Net v2	No	CBCT	Detect	No visual explainability; restricted insight into model decisions
Proposed	Hybrid (CNN + Transformer)	EfficientNetB0 + Swin Transformer + Grad-CAM	Yes	CT images	Classify	----

3. Proposed Methodology

In this study, a hybrid deep learning model is proposed for the classification of maxillary sinus abnormalities from CT images, leveraging both convolutional neural networks (CNN) and Swin Transformers to capture local and global features, respectively, as shown in Figure 1. The dataset consists of balanced medical CT images, categorized into four classes: Normal MS, Opacified MS, Polyposis, and Retention Cysts. Data preprocessing and augmentation techniques, such as resizing, histogram equalization, orientation stan-

dardization, and class balancing, are applied to enhance the robustness and diversity of the dataset. The model architecture includes EfficientNetB0 as a feature extractor for low- to mid-level features and Swin Transformer for capturing long-range dependencies, followed by attention-based fusion to combine the outputs of both models. For 4 class classification, a custom classification head is employed. Cross entropy loss and Adam optimizer are used for model training with learning rate tuning using Optuna. It monitors accuracy and loss for training process and employs early stopping to avoid overfitting. Optuna is used to optimize hyperparameters using Bayesian optimization so as to efficiently explore the parameter space. To aid the model's transparency and interpretability, Gradient weighted Class Activation Mapping (Grad-CAM) is then used to provide visualizations of the areas in the image that contributed most to the model's decision.

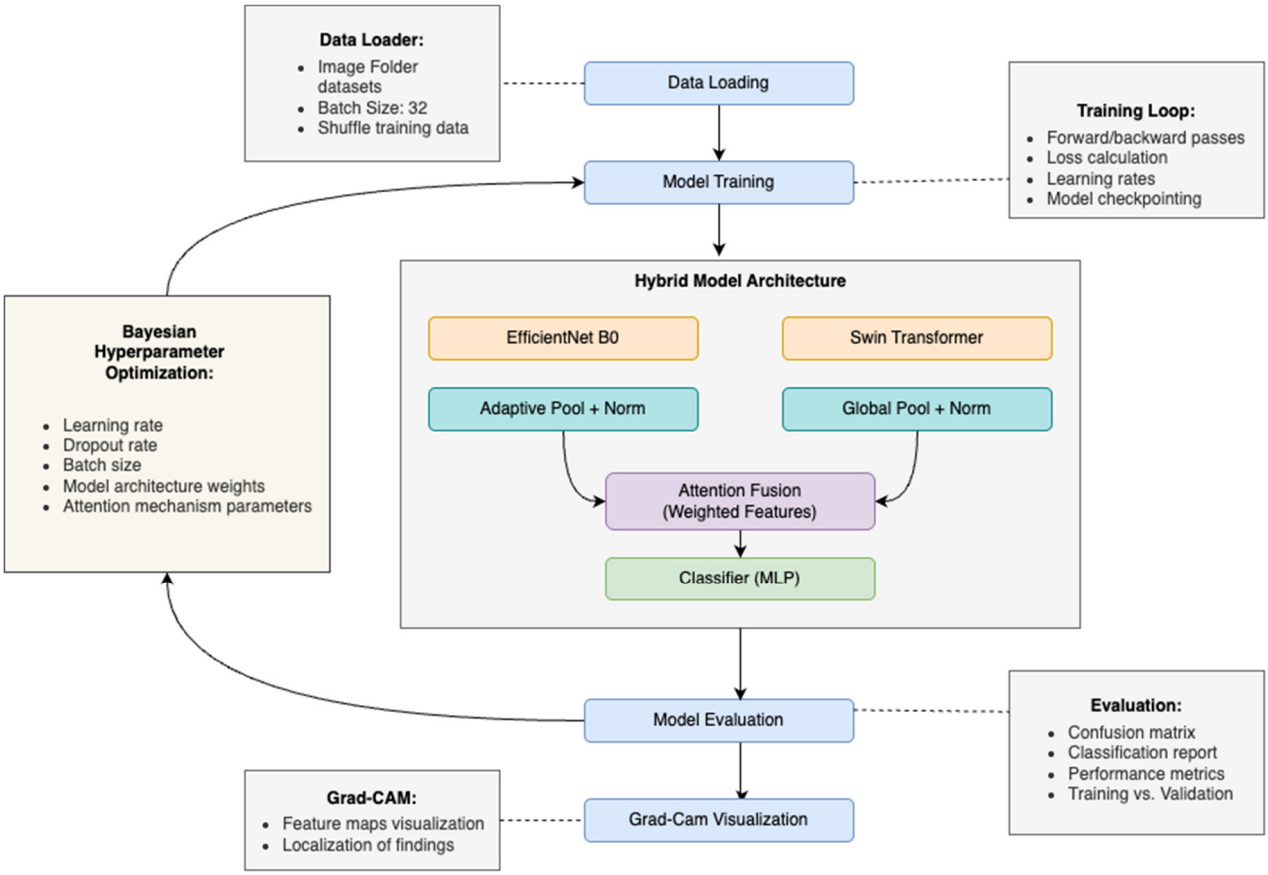


Figure 1. Architecture of the Proposed Model.

3.1. Dataset Description

The dataset stands as a fundamental requirement for assessing and developing the proposed maxillary sinus classification model. The dataset comprises CT images at high resolution which are collected from different healthcare institutions for evaluations on symptomatic patients. The data preparation process focused on maintaining both quality and consistency while ensuring uniformity between image acquisition procedures. Expert radiologists performed image classification labeling to confirm proper assignment of different classes. The dataset was preprocessed and augmented in several ways in order for the model to be effectively trained and evaluated as shown in Figure 2.

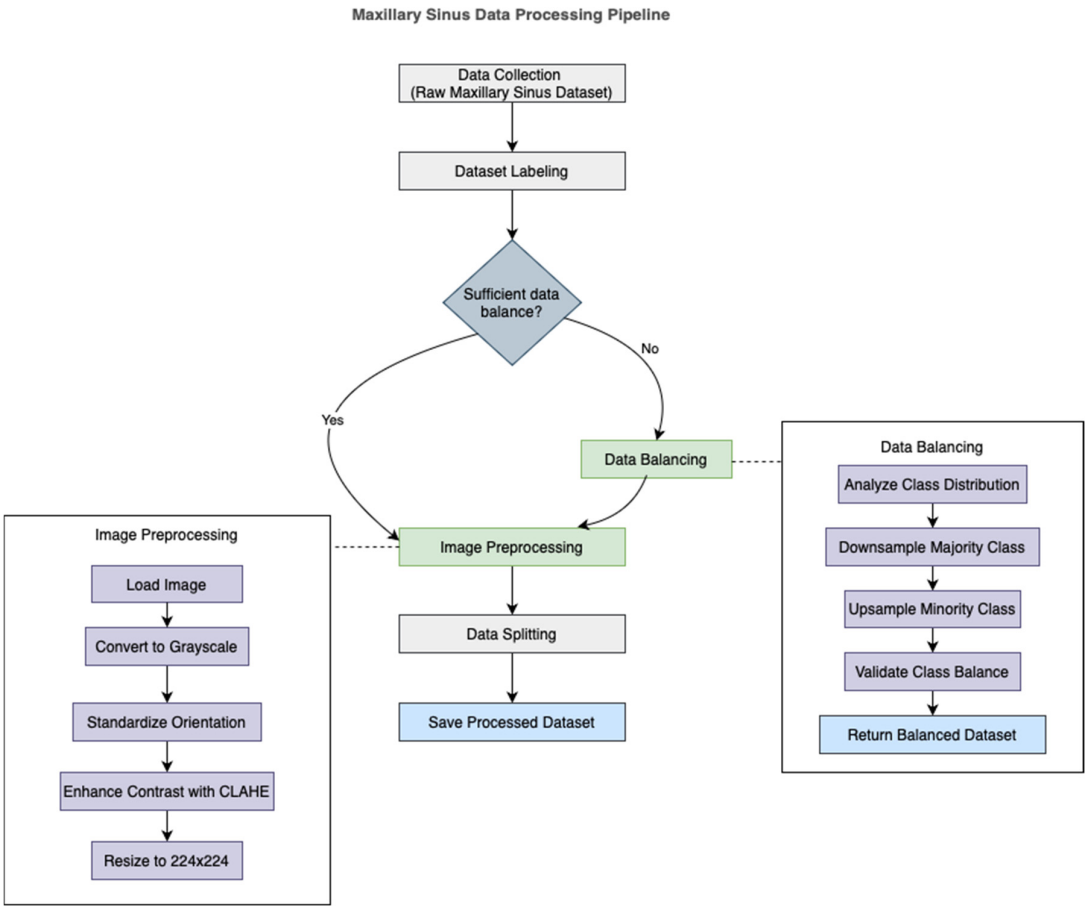


Figure 2. Flowchart of the Data Processing Pipeline.

3.1.1. Data Collection

The CT data for symptomatic and referred patients was collected from two healthcare institutions based in Hail, Saudi Arabia, during the years 2022 to 2024, with all images anonymized to preserve patient identity and confidentiality in accordance with standard medical imaging research protocols. The Institutional Review Board (IRB) of Ministry of Health, Hail, Saudi Arabia approved this dataset (https://www.moh.gov.sa/en/Pages/Default.aspx, accessed on 26 December 2024) [37]. The dataset included patients who belonged to multiple demographics ranging from both genders and various age groups. All selected patients were at least 18 years old because sinusitis affects adults more commonly [4]. To obtain consistent anatomical features, the data were restricted to adults to avoid variations in sinus size and paranasal sinuses' pneumatization due to age.

The CT scans used in this study were limited to coronal views and were performed without using contrast agents. The chosen images possessed 0.2 mm slice thickness providing high-resolution necessary for maxillary sinus evaluation. The 0.2 mm thick sliced images provide optimal capabilities for identifying any opacifications within the maxillary sinus region. CT image selection followed established criteria to guarantee dataset reliability and consistency. Exclusion criteria were designed to exclude patients with anatomical variations or disease, which may alter the normal or pathological appearance of the maxillary sinus. For consistency in imaging protocols and to concentrate on relevant soft tissue features, strict inclusion criteria were also set. The detailed criteria are summarized in Table 2.

Table 2. Inclusion and exclusion criteria for CT image selection.

Category	Criteria
Exclusion Criteria	No congenital anomalies
	No history of trauma
	No previous surgeries
	No history of drug use or smoking
Inclusion Criteria	Coronal view CT scans
	Soft tissue window
	CT scans without contrast

3.1.2. Data Quality Assessment and Expert Labeling

Once the CT image data was collected, a preliminary analysis was conducted to assess data quality and to select appropriate samples for further investigation. Images exhibiting noise artifacts, poor resolution, or missing information were excluded from the dataset in the first instance. This quality control step refined the dataset and ensured that only diagnostically reliable images were included for subsequent processing.

Following quality control, two experienced radiologists were engaged to manually evaluate and label the maxillary sinus findings into four distinct classes: Normal Maxillary Sinus (Normal MS), Opacified Maxillary Sinus (Opacified MS), Polyposis, and Retention Cysts. A visual illustration of the classification is presented in Figure 3, The first column (form left) column shows a normal sinus, the second column depicts partial opacification (Opacified MS), and the third and fourth columns represent Polyposis and Retention Cysts, respectively.

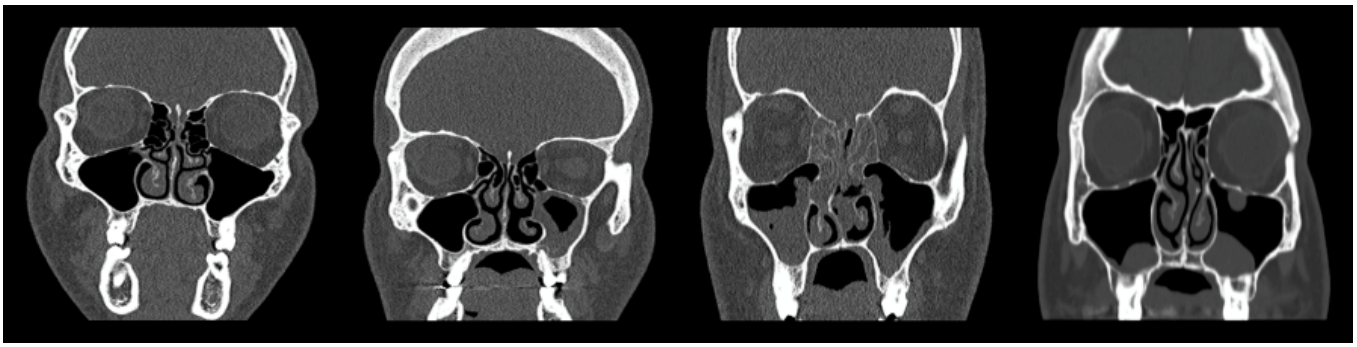


Figure 3. CT samples illustrating representative cases from each diagnostic category.

Recognizing that visual interpretation of CT images can be subjective, and different radiologists may interpret images differently based on their experience, interobserver variability was expected. To formally measure the level of agreement, Cohen's Kappa [38] was employed, resulting in a substantial agreement (k = 0.821). Based on mutual consensus, the collected images were reliably categorized into the four predefined classes.

3.2. Preprocessing

The preprocessing stage played a critical role in standardizing input data and enhancing feature visibility, which are essential for effective model training. These steps ensured consistency, reduced data bias, and improved the reliability of the classification framework.

3.2.1. Image Standardization and Enhancement Procedures

All CT images were resized to 224 $\times$ 224 pixels to ensure compatibility with both EfficientNetB0 and Swin Transformer architectures. Additionally, pixel intensity values were

normalized to the [0, 1] range to standardize input data and facilitate efficient model training.

Given that lighting inconsistencies are common in medical imaging, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to enhance image contrast. This technique improves the visibility of features, especially in low-contrast regions. The parameter settings for CLAHE are summarized in Table 3.

Table 3. CLAHE parameter values for contrast enhancement.

Parameter	Value	Explanation
Clip Limit	2.0	Limits the contrast amplification
Tile Grid Size	(8, 8)	Divides the image into 8 × 8 tiles for local enhancement

To maintain consistency across the dataset, Otsu's thresholding was applied for image binarization, followed by contour-based rotation correction. This ensured that all CT images were correctly oriented, a critical requirement for reliable medical image analysis.

3.2.2. Dataset Balancing

The original dataset exhibited class imbalance, with Normal MS being overrepresented compared to the other categories. To address this, the dataset was balanced by employing the following strategies:

- Downsampling: The Normal MS class was reduced to 400 samples by selecting images based on quality criteria such as clarity, anatomical completeness, and absence of motion artifacts.
- Upsampling: The minority classes (Opacified MS, Polyposis, Retention Cysts) were augmented to reach 400 samples each using image augmentation techniques.

As a result, the final dataset consisted of 1600 images, evenly distributed across the four classes: Normal Maxillary Sinus (Normal MS), Opacified MS, Polyposis, and Retention Cysts. The images were labeled with the assistance of medical experts specialized in maxillary sinus evaluation and organized into Train, Validation, and Test sets. The final distribution of images across classes is summarized in Table 4.

Table 4. Dataset distribution across different maxillary sinus conditions.

Class	Initial Image Count	Final Image Count
Normal MS	772	400
Opacified MS	203	400
Polyposis	198	400
Retention Cysts	201	400

3.2.3. Data Augmentation

Data augmentation was implemented to enhance model robustness and generalization. Variations were introduced in the training images through a combination of random transformations. The augmentation techniques and their respective parameter settings are detailed in Table 5.

Table 5. Summary of data augmentation parameters.

Augmentation Type	Parameter Values	Explanation
Rotation	±10°	Random rotation within ±10° to simulate angle variations.
Zoom	0.8 to 1.2	Random zoom to simulate varying object distances.
Shift	±0.2 of width/height	Random shift of up to 20% of image dimensions.
Shear	0.2	Random shear to simulate distortion in image.
Brightness	0.7 to 1.3	Random adjustment of image brightness.
Horizontal Flip	50%	Flip the image horizontally with a 50% chance.

After balancing, the dataset was divided into training, validation, and test sets following a 70:15:15 ratio as shown in Table 6.

Table 6. Data Splitting Details.

Set	% of Total Dataset	Purpose
Training Set	70%	Used for model learning and training.
Validation Set	15%	Used for fine-tuning hyperparameters and monitoring model performance during training.
Test Set	15%	Reserved exclusively for final model evaluation to provide unbiased performance metrics.

3.3. Hybrid Model Architecture

The proposed classification framework utilizes a hybrid model architecture that integrates convolutional and transformer-based feature extraction methods. This design leverages the strengths of both EfficientNetB0 and Swin Transformer to extract comprehensive local and global features from the maxillary sinus CT scans. The overall architecture, as shown in Figure 1, includes three main components: backbone feature extractors, attention-based feature fusion, and a custom classification head.

3.3.1. Backbone Feature Extractors

EfficientNetB0, a convolutional neural network optimized using compound scaling, serves as one of the primary feature extractors [39]. Pre-trained on ImageNet, it offers a strong starting point for transfer learning and is particularly effective at capturing low-level spatial textures and mid-level semantic features, which are essential for medical image analysis. The reduced number of parameters enables the network to converge more quickly throughout the training period. The feature expressiveness of EfficientNetB0 benefits from depthwise separable convolutions and squeeze-and-excitation optimization blocks that maintain computational efficiency. The second backbone is the Swin Transformer, a hierarchical vision transformer model that applies self-attention mechanisms within shifted windows [40]. This model excels at capturing long-range dependencies and understanding complex anatomical structures in high-resolution CT images. The pyramid architecture of Swin Transformer enables multi-scale features that match the localized extraction per-

formed by EfficientNetB0. Through its shifted windowing scheme, the Swin Transformer maintains global context modeling alongside efficient large image processing.

3.3.2. Attention-Based Feature Fusion

The outputs from both EfficientNetB0 and Swin Transformer are first normalized to ensure consistency in feature scale. L2 normalization scales each feature vector such that the sum of the squares of its components equals one, effectively projecting it onto the unit hypersphere. This helps in reducing the influence of feature magnitude differences and emphasizes the directionality of features, which is especially important in attention-based fusion. L2 normalization ensures uniform feature magnitude, stabilizes training, improves convergence, and prevents the model from becoming biased toward features with larger scales. These feature maps are then fused using learnable attention weights:

$$F_{fused} = w_1.F_{effNet} + w_2.F_{swin},\ Where\ w_1 + w_2 = 1 \quad (1)$$

This equation represents the weighted feature fusion where the model learns optimal attention weights w_1 and w_1 during training. Through learnable attention weight assignment, the model dynamically focuses, respectively, on the most important backbone features in each input image which enables accurate contextual decision output. The combined features help the model adjust its focus according to image properties which leads to improved generalization. Additionally, this fusion strategy enables the model to combine both local and global context effectively, leveraging EfficientNetB0's strength in capturing detailed local patterns and Swin Transformer's capability to model long-range dependencies. The attention-based fusion module details are given in Table 7.

Table 7. Attention-Based Fusion Module Details.

Component	Implementation	Output Dimension	Notes/Changes
Input Features	EfficientNetB0 (1280-dim), Swin Transformer (768-dim)	1280, 768	Backbone outputs, normalized before fusion
Normalization	LayerNorm (1280 for EffNet, 768 for Swin)	1280, 768	Change from previously described L2 norm
Concatenation	Concatenate normalized features	2048	Prepares for attention weight computation
Attention Weight Computation	Linear (2048 $\rightarrow$ 2) + Softmax	2	Generates sample-specific weights [w_1, w_2]
Weighted Scaling	Multiply normalized features by corresponding attention weights	1280, 768	Feature vectors scaled per sample
Fusion	Concatenate scaled features	2048	Final fused representation for classifier
Key Benefit	Adaptive focus on local and global features	2048	Improves generalization and context-aware decision making

3.3.3. Custom Classification Head

The fused feature vector is passed through a custom classification head composed of a fully connected layer, a dropout layer, and a final softmax layer for four-class classification. The architecture is defined as:

$$\hat{y} = softmax\Big(W_2\Big(Dropout\Big(ReLU\Big(W_1 F_{fused} + b_1\Big)\Big) + b_2\Big) \quad (2)$$

where W_1 and W_2 are learned weight matrices, b_1 and b_2 are biases, and $\hat{y}$ is the predicted probability distribution. The ReLU activation function introduces non-linearity to the

network, allowing it to learn complex patterns and relationships in the data, while the Dropout layer helps prevent overfitting by randomly deactivating a fraction of neurons during training. The softmax function at the output layer ensures that the model produces a probability distribution over the four classes, making it suitable for multi-class classification tasks.

The best configuration from hyperparameter tuning adopted 256 dense units accompanied by a dropout rate of 0.4. Model complexity and generalization were balanced by selecting the dropout rate while the number of dense units was chosen to achieve sufficient learning capacity without overfitting. Backpropagation with the Adam optimizer updates the model weights during training by automatically adjusting learning rates for faster convergence and improved training speed.

3.4. Model Training

3.4.1. Loss Function

The training objective is to minimize the standard cross-entropy loss for multi-class classification:

$$L_{CE} = -\sum_{i*1}^{C} y_i \log(\hat{y}_i) \quad (3)$$

where y_i is the ground truth label and $\hat{y}_i$ is the predicted probability for class i. Although weighted cross-entropy was also tested to account for potential imbalance, it did not show significant improvement over the standard loss function.

3.4.2. Optimizer and Learning Rate Scheduling

The Adam optimizer is utilized with a weight decay of 0.01 to prevent overfitting. The initial learning rate, optimized using Bayesian methods, is set at 0.0003. During training, StepLR and ReduceLROnPlateau schedulers are applied to adjust the learning rate based on the plateauing of validation loss. Training achieved a high accuracy of 98.6% on the training set and 96.4% on the validation set. Early stopping was employed to halt training when no improvement in validation loss was observed, thereby preventing overfitting.

3.4.3. Training Monitoring

Throughout training, model performance was continuously tracked. The model was checkpointed whenever a new highest validation accuracy was recorded. Loss and accuracy trends for both training and validation sets were visualized post-training to ensure model convergence and stability.

3.5. Hyperparameter Optimization

3.5.1. Optimization Strategy

Optuna, a hyperparameter optimization framework based on Bayesian optimization using Tree-structured Parzen Estimators (TPE), was employed for tuning [41]. It offers an efficient search strategy that prunes non-promising trials early, thus saving computational resources. The tuned parameters values are shown in Table 8.

Table 8. Tuned Parameters and Best Values.

Parameter	Range/Options	Best Value
Learning Rate	Log-uniform (1×10^{-5} to 1×10^{-3})	0.0003
Batch Size	[16, 32, 64]	32
Dropout	Uniform (0.1 to 0.5)	0.4
Dense Units	[128, 256, 512]	256
Epochs	Fixed during tuning [10, 20, 30]	13

3.5.2. Objective Function and Outcome

The custom objective function trains the model for a fixed number of epochs and evaluates its performance using the validation accuracy from the final epoch. Early stopping and pruning mechanisms are incorporated to terminate unproductive trials. Hyperparameter tuning led to a 2% increase in test accuracy compared to the baseline. Before tuning, the model achieved 92.92% test accuracy, which improved to 95.83% after optimization.

4. Results and Discussion

The performance of the model on the test dataset is evaluated by standard evaluation metrics like accuracy, loss, precision, recall, and *F*1 *score* and the evaluation of each metric is presented in this section. A general measure of a model's performance is the accuracy, which is calculated as the proportion of correctly classified instances in the whole dataset. Precision is the ability of the model to minimize false positives; it is the ratio of correctly identified positives to the total number of positives that were predicted. Recall is the sensitivity of the model, i.e., the ratio of correctly identified actual positive cases to the total number of actual positive cases. *F*1 *score* is the harmonic mean of precision and recall that gives a balanced view of whether the model can avoid both false positives and false negatives. In medical imaging, where incorrect diagnosis can be detrimental, the *F*1 *score* is especially useful. Below are the formal definitions of these metrics as a means of statistical clarity to evaluate classification results: *TP* = true positive, *FP* = false positive, *TN* = true negative, and *FN* = false negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

4.1. Confusion Matrix Analysis

The confusion matrix (Figure 4) provides a detailed breakdown of the model's class-wise prediction outcomes. It reveals that the model performed exceptionally well in distinguishing Normal Maxillary Sinus (MS) and Retention Cysts, correctly identifying 59 out of 60 and 58 out of 60 instances, respectively. However, minor misclassifications occurred between the Opacified MS and Polyposis classes, which can be attributed to the overlapping radiological characteristics of fluid retention and polypoidal soft tissue densities. For instance, four Polyposis cases were incorrectly labeled as Retention Cysts.

This pattern of misclassification reflects real-world diagnostic challenges faced by radiologists, where CT findings of certain pathological categories may present similarly. Hence, the model's errors are clinically plausible and support its potential for assisting diagnostic workflows rather than contradicting clinical expertise.

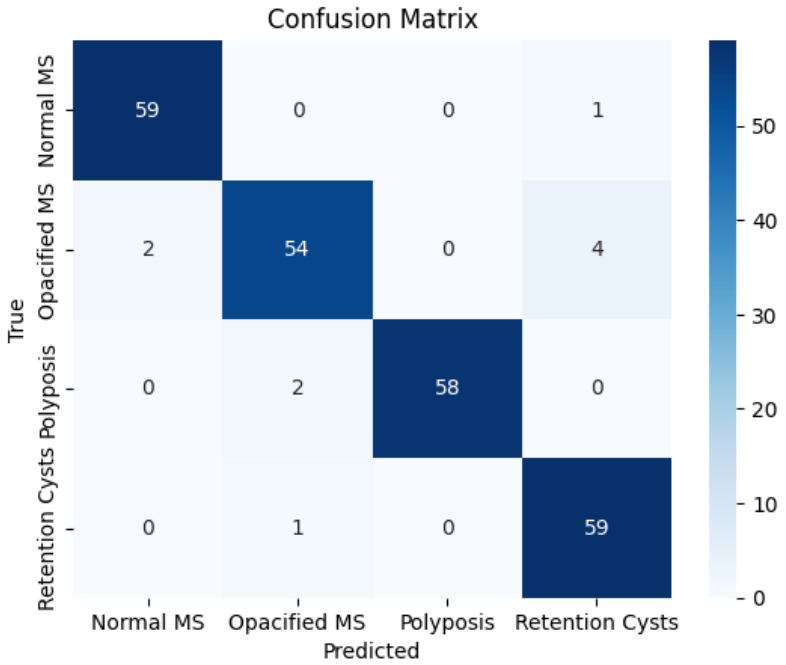


Figure 4. Confusion of the Metrics of the Proposed Model.

4.2. Classification Performance Metrics

Standard metrics like precision, recall, *F1-score* and accuracy were used to evaluate the classification performance. Additionally, we calculated these metrics per class as well as aggregated for a macro-level model performance measure. Overall accuracy on all categories was 95.83%, showing strong generalization for all categories. Opacified MS and Normal MS precision values ranged from 0.92 to 0.98, implying a high capacity to minimize false positives. Similar recall values were calculated with a minimum of 0.92 for Opacified MS and 0.98 for Normal MS, suggesting that the model is able to detect true positives accordingly. This ranged from 0.92 to 0.98 *F1-score*, indicating consistent and reliable results across all four classes. Figure 5 summarizes the detailed metrics.

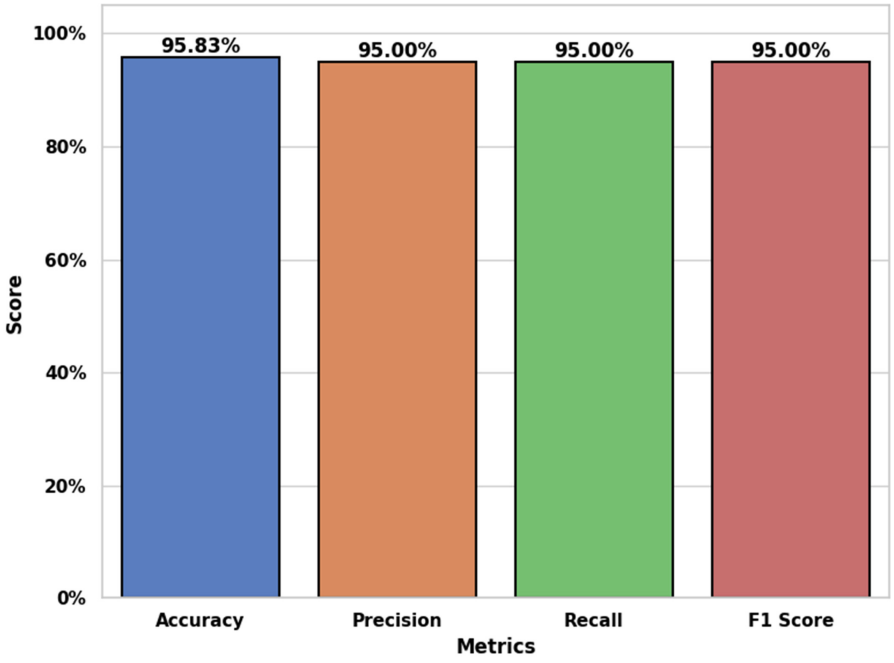


Figure 5. Model Performance Metrics.

4.3. ROC-AUC Analysis

The ROC-AUC was computed to further evaluate the discriminative ability of the model as shown in Figure 6. It yielded an ROC-AUC score of 0.982, which separated classes extremely well. The class variable wise AUC values were Normal MS = 1.00, Opacified MS = 0.98, Polyposis = 1.00 and Retention Cysts = 0.99. These values confirm the ability of the model to reliably differentiate healthy from diseased sinuses, as well as more subtle pathological examples.

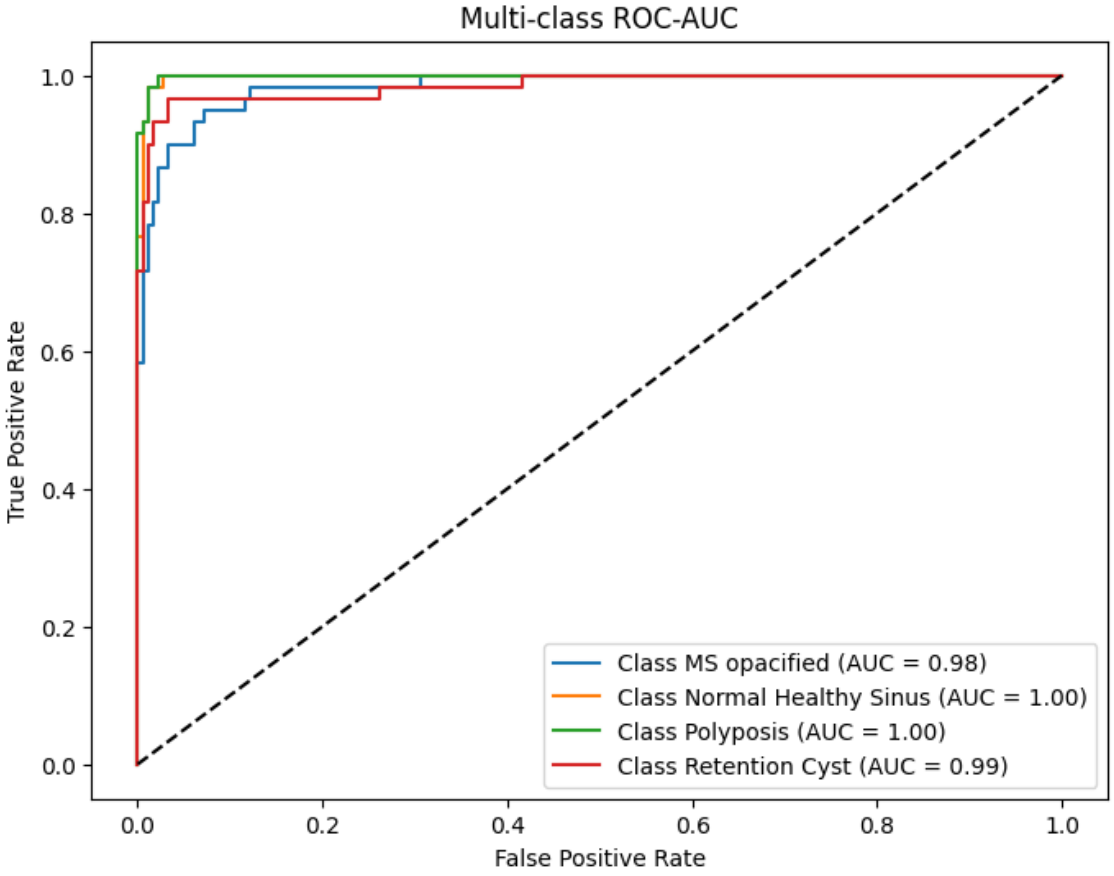


Figure 6. ROC-AUC Curve of the Proposed Model.

The ROC-AUC performance also reinforces the effectiveness of the hybrid architecture in combining local and global features to optimize classification, especially for classes with overlapping imaging patterns.

4.4. Training Dynamics and Model Convergence

Training dynamics were monitored closely to assess convergence behavior and detect potential overfitting. The training and validation loss curves (Figure 7) exhibited stable and monotonic decline, with the training loss reaching a minimum of 0.2 and validation loss stabilizing around 0.4. Correspondingly, the model achieved peak training accuracy of 98.6% and validation accuracy of 96.4%.

To prevent overfitting, early stopping was employed, terminating training once the validation loss plateaued. Model checkpoints ensured that the best-performing model was preserved for evaluation. The consistent trend across training and validation curves indicates a well-regularized model with strong generalization capability.

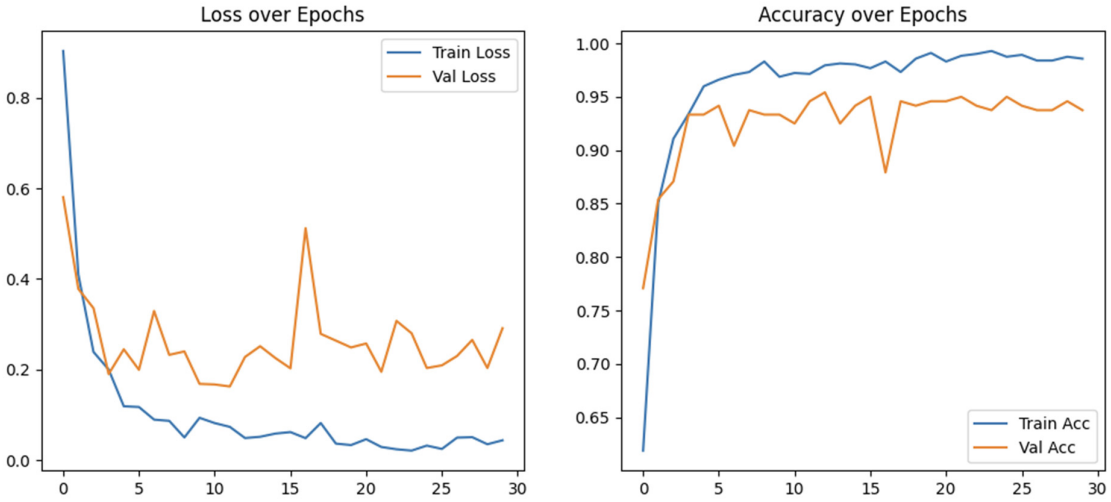


Figure 7. Visualization of training and validation accuracy and loss.

4.5. Comparison with Related Studies

To the best of our knowledge, no prior work has directly addressed the automated four-class classification of maxillary sinus abnormalities from CT images using a hybrid deep learning framework. However, some studies within the broader domain of sinus disease detection have utilized MRI, X-ray, and panoramic radiographs. For example, Bhattacharya et al. [26] employed MRI for binary classification of sinusitis, Lim et al. [18] analyzed sinus opacification using X-ray images, and Murata et al. [31] investigated panoramic radiographs for detecting maxillary sinus lesions. While these studies provide useful insights, they differ significantly in imaging modality, classification scope, and explainability. Our study contributes uniquely by combining CNN and Transformer architectures with an attention-based fusion mechanism and Grad-CAM visualization, achieving high diagnostic accuracy while maintaining transparency for clinical use. A detailed comparison is summarized in Table 9.

Table 9. Comparison of our proposed hybrid deep learning framework with related studies.

Study	Imaging Modality	Task/Classes	Methodology	Performance Metrics	Explainability	Limitation Compared to Our Work
Bhattacharya et al. [26]	MRI	Binary (Sinusitis vs. Normal)	CNN-based model	Accuracy: ~87%, Precision: NR, Recall: NR, F1: NR, AUC: NR	No	Limited to binary classification, lacks interpretability
Lim et al. [18]	X-ray	Sinus opacification (Binary/Partial detection)	Deep learning on 2D radiographs	Accuracy: 80–85%, Precision: NR, Recall: NR, F1: NR, AUC: NR	No	Lower sensitivity, non-gold standard imaging
Murata et al. [31]	Panoramic Radiographs	Maxillary sinus lesions (Binary)	Conventional ML + handcrafted features	Accuracy: ~82%, Precision: NR, Recall: NR, F1: NR, AUC: NR	No	Non-CT modality, limited diagnostic value
Our Study	CT (Gold Standard)	Four-class (Normal, Opacified, Polyposis, Retention Cysts)	Hybrid EfficientNetB0 + Swin Transformer with Attention Fusion	Accuracy: 95.83%, Precision: 0.95, Recall: 0.95, F1: 0.95, AUC: >0.98	Yes (Grad-CAM)	First to combine hybrid DL + CT + explainability for sinus classification

4.6. Interpretation and Clinical Relevance

The hybrid model's success can be attributed to the complementary strengths of its constituent backbones. EfficientNetB0 contributed robust spatial and texture-level feature extraction, while the Swin Transformer provided the capacity to model long-range

dependencies across the input volume. Their integration via attention-based fusion enabled adaptive weighting of both feature types, enhancing the model's discriminative power.

Visualizations of Grad-CAMs confirmed model attention was appropriately localized to anatomically meaningful regions, such as sinuses walls, opacified cavities, and mucosal thickening. The model's interpretability is further strengthened by these visual explanations that are key for clinical adoption. Moreover, the model also showed resilience to class imbalance as a result of the augmentation strategies and class rebalancing techniques used during preprocessing. Consequently, it exhibited minimal degradation in precision and recall across these minority classes (Opacified MS, Polyposis).

5. Grad-CAM Visualization

Interpretability is an important aspect of building trust with artificial intelligence systems in medical imaging. To address this, Gradient weighted Class Activation Mapping (Grad-CAM) [42] was used to visualize and interpret the predictions made by the proposed hybrid deep learning model. It gives visual explanations by highlighting the regions in an image that most contributed to the model's predictions using Grad-CAM. The saliency maps bridge the gap between algorithmic output and clinical understanding of maxillary sinus conditions, and can be used as a tool for validation that the model is classifying the maxillary sinus condition using anatomically relevant structures.

5.1. Methodological Framework of Grad-CAM

The Grad-CAM technique was implemented on the EfficientNetB0 backbone of the hybrid architecture. The process begins by calculating the gradients of the output class score with respect to the feature maps from the last convolutional layer. These gradients are globally averaged to determine the relative importance of each feature map channel for the predicted class. The resulting importance weights are then combined with the feature maps, followed by the application of a ReLU function to retain only positively contributing activations. This produces a coarse localization map, or heatmap, which is subsequently upsampled to the input image resolution (224×224 pixels) and normalized.

Finally, the heatmap is superimposed on the original CT image using a color gradient, typically ranging from blue (low importance) to red (high importance). This overlay visually indicates where the model is attending when making its prediction, thereby enhancing interpretability.

5.2. Implementation and Representative Sampling

For consistency and diagnostic relevance, the final convolutional layer of the EfficientNetB0 network was selected as the target layer due to its high semantic content and retained spatial resolution. Representative CT images from all four classes—Normal Maxillary Sinus (MS), Opacified MS, Polyposis, and Retention Cysts—were selected for Grad-CAM analysis. Post-processing techniques, including heatmap thresholding, were applied to reduce visual noise and to sharpen the focus on medically relevant anatomical structures.

In Retention Cysts, the heatmaps displayed sharply localized activations along cyst margins, while Normal MS scans exhibited diffused, low-intensity activations across the sinus region, accurately reflecting the absence of pathology. Figure 8 presents the Grad-CAM heatmaps for Retention Cysts cases.

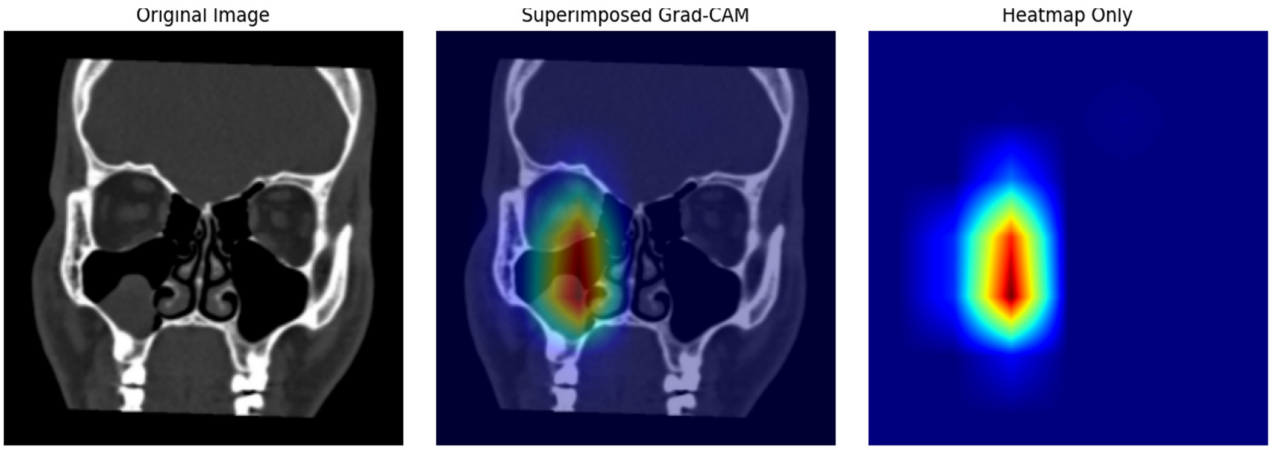


Figure 8. Grad-CAM heat maps Retention Cysts.

Figure 9 shows the resulting heatmaps for pathological cases, capturing variations in activation intensity and localization patterns.

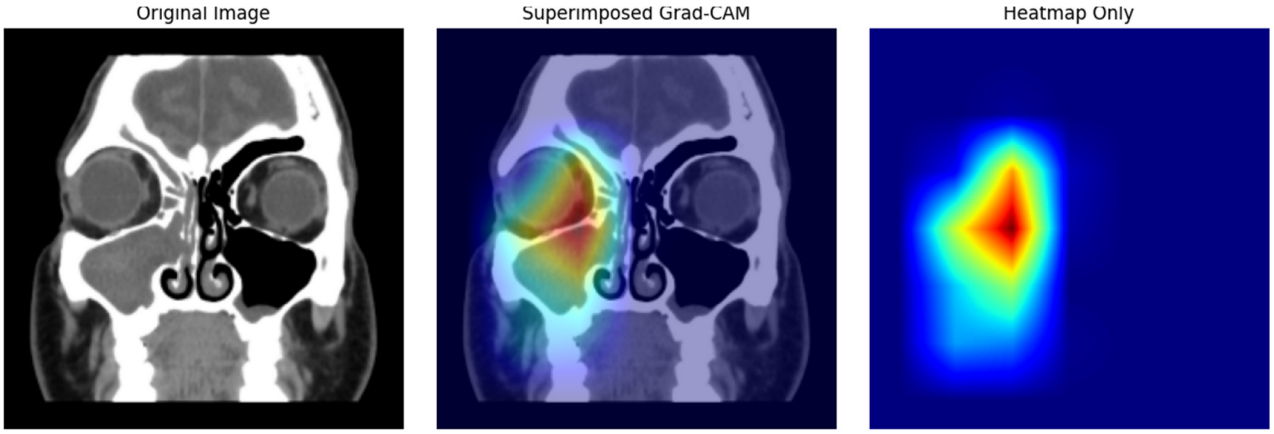


Figure 9. Grad-CAM heat maps Polyposis.

5.3. Interpretation

Analysis of the Grad-CAM outputs revealed strong alignment between the model's focus areas and clinically significant regions. In pathological cases such as Opacified MS and Polyposis, the model demonstrated high-intensity activations within the sinus cavities, particularly around regions exhibiting mucosal thickening or fluid accumulation. These activations corresponded closely with radiological markers typically used for diagnosis.

Grad-CAM was also utilized for error analysis. Misclassifications, particularly between Opacified MS and Polyposis, were associated with overlapping activation patterns in the heatmaps.

6. Conclusions and Future Work

In conclusion, this study presented a robust hybrid deep learning framework combining EfficientNetB0 and Swin Transformer architectures for the automated classification of maxillary sinus from CT images. The model achieved exceptional performance, with 95.83% test accuracy and strong discriminative capability across all classes, as evidenced by ROC-AUC scores exceeding 0.98. Importantly, we demonstrated the efficacy of Grad-CAM visualizations in elucidating the model's decision making process, revealing its alignment with clinically relevant anatomical features. These explainability insights not only validate the model's reliability but also enhance confidence among medical practitioners, a critical

factor for clinical adoption. The proposed framework offers a promising tool for accurate, interpretable, and clinically relevant sinusitis diagnosis, paving the way for broader integration of AI-assisted imaging in routine medical practice.

Despite these promising results, the study's limitations highlight key areas for future improvement. The primary constraint remains the relatively small size of the clinical dataset, which, despite augmentation techniques, may limit the model's generalization to rare or complex cases. To address this, future work should focus on expanding the dataset through multi-institutional collaborations, incorporating diverse demographic and pathological variations. Additionally, integrating multi-view CT data (e.g., sagittal and axial planes) could enhance the model's spatial understanding of sinus structures, potentially improving diagnostic precision for conditions with subtle radiographic differences.

Author Contributions: Conceptualization, M.A. and A.G.F.; methodology, M.A. and A.G.F.; software, M.A.; validation, M.A. and A.G.F.; investigation, M.A. and A.G.F.; resources, M.A.; data curation, M.A. and A.G.F.; writing—original draft preparation, M.A. and A.G.F.; writing—review and editing, M.A.; visualization, M.A.; supervision, A.G.F. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Hail Health Cluster, Hail, Saudi Arabia, with number H-08-L-074-2023-72.

Informed Consent Statement: Patient consent was waived by the IRBs because of the retrospective nature of this investigation and the use of anonymized patient data.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request, subject to the approval of the Institutional Review Boards of the participating institutions.

Acknowledgments: The authors gratefully acknowledge the support provided by the Faculty of Computing and Information Technology (FCIT), King Abdulaziz University (KAU), Jeddah, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Min, H.K.; Lee, S.; Kim, S.; Son, Y.; Park, J.; Kim, H.J.; Lee, J.; Lee, H.; Smith, L.; Rahmati, M.; et al. Global Incidence and Prevalence of Chronic Rhinosinusitis: A Systematic Review. *Clin. Exp. Allergy* **2025**, *55*, 52–66. [CrossRef]
2. Toppila-Salmi, S.K. *European Forum for Research and Education in Allergy and Airway Diseases (EUFOREA)*; University of Helsinki: Helsinki, Finland, 2017.
3. Battisti, A.S.; Modi, P.; Pangia, J. *Sinusitis (Archived)*; StatPearls: St. Petersburg, FL, USA, 2023.
4. Alotaibi, A.D.; Zafar, M.; Alsuwayt, B.N.; Raghib, R.N.; Elhaj, A.H. Body Mass Index and Related Risk Factor of Sinusitis Among Adults in Saudi Arabia: A Cross-Sectional Study. *Cureus* **2023**, *15*, e40454. [CrossRef]
5. Whyte, A.; Boeddinghaus, R. The maxillary sinus: Physiology, development and imaging anatomy. *Dentomaxillofac. Radiol.* **2019**, *48*, 20190205. [CrossRef]
6. Ketabchi, A.; Ahmed, N. Orofacial infections. In *Maxillofacial Surgery*; Churchill Livingstone: Amsterdam, The Netherlands, 2017.
7. Aaløkken, T.M.; Hagtvedt, T.; Dalen, I.; Kolbenstvedt, A. Conventional sinus radiography compared with CT in the diagnosis of acute sinusitis. *Dentomaxillofac. Radiol.* **2003**, *32*, 60–62. [CrossRef]
8. Gregurić, T.; Prokopakis, E.; Vlastos, I.; Doulaptsi, M.; Cingi, C.; Košec, A.; Zadravec, D.; Kalogjera, L. Imaging in chronic rhinosinusitis: A systematic review of MRI and CT diagnostic accuracy and reliability in severity staging. *J. Neuroradiol.* **2021**, *48*, 277–281. [CrossRef]
9. Kandukuri, R.; Phatak, S. Evaluation of Sinonasal Diseases by Computed Tomography. *J. Clin. Diagn. Res.* **2016**, *10*, TC09. [CrossRef]

10. Neagos, A.; Dumitru, M.; Vrinceanu, D.; Costache, A.; Marinescu, A.N.; Cergan, R. Ultrasonography used in the diagnosis of chronic rhinosinusitis: From experimental imaging to clinical practice. *Exp. Ther. Med.* **2021**, *21*, 611. [CrossRef]
11. Leonard, S.; Sinha, A.; Reiter, A.; Ishii, M.; Gallia, G.L.; Taylor, R.H.; Hager, G.D. Evaluation and Stability Analysis of Video-Based Navigation System for Functional Endoscopic Sinus Surgery on In Vivo Clinical Data. *IEEE Trans. Med. Imaging* **2018**, *37*, 2185–2195. [CrossRef]
12. Stenner, M.; Rudack, C. Diseases of the nose and paranasal sinuses in child. *GMS Curr. Top. Otorhinolaryngol. Head Neck Surg.* **2014**, *13*, Doc10. [CrossRef]
13. Ziegler, A.; Patadia, M.; Stankiewicz, J. Neurological complications of acute and chronic sinusitis. *Curr. Neurol. Neurosci. Rep.* **2018**, *18*, 5. [CrossRef]
14. Mayerhoefer, M.E.; Materka, A.; Langs, G.; Häggström, I.; Szczypiński, P.; Gibbs, P.; Cook, G. Introduction to Radiomics. *J. Nucl. Med.* **2020**, *61*, 488–495. [CrossRef]
15. Aurelia, J.E.; Rustam, Z.; Laeli, A.R.; Maulidina, F. Neural Network-Support Vector Machine for Sinusitis Classification. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020; pp. 1185–1189. [CrossRef]
16. Barragán-Montero, A.; Javaid, U.; Valdés, G.; Nguyen, D.; Desbordes, P.; Macq, B.; Willems, S.; Vandewinckele, L.; Holmström, M.; Löfman, F.; et al. Artificial intelligence and machine learning for medical imaging: A technology review. *Phys. Medica* **2021**, *83*, 242–256. [CrossRef]
17. Kim, H.G.; Lee, K.M.; Kim, E.J.; Lee, J.S. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models. *Quant. Imaging Med. Surg.* **2019**, *9*, 942–951. [CrossRef]
18. Lim, S.H.; Kim, J.H.; Kim, Y.J.; Cho, M.Y.; Jung, J.U.; Ha, R.; Jung, J.H.; Kim, S.T.; Kim, K.G. Aux-MVNet: Auxiliary Classifier-Based Multi-View Convolutional Neural Network for Maxillary Sinusitis Diagnosis on Paranasal Sinuses View. *Diagnostics* **2022**, *12*, 736. [CrossRef]
19. Laura, C.O.; Hofmann, P.; Drechsler, K.; Wesarg, S. Automatic detection of the nasal cavities and paranasal sinuses using deep neural networks. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1154–1157. [CrossRef]
20. Afshar, P.; Mohammadi, A.; Plataniotis, K.N.; Oikonomou, A.; Benali, H. From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities. *IEEE Signal Process. Mag.* **2019**, *36*, 132–160. [CrossRef]
21. Hesamian, M.H.; Jia, W.; He, X.; Kennedy, P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J. Digit. Imaging* **2019**, *32*, 582–596. [CrossRef]
22. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
23. Jeon, Y.; Lee, K.; Sunwoo, L.; Choi, D.; Oh, D.Y.; Lee, K.J.; Kim, Y.; Kim, J.W.; Cho, S.J.; Baik, S.H.; et al. Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs. *Diagnostics* **2021**, *11*, 250. [CrossRef]
24. Kuwana, R.; Ariji, Y.; Fukuda, M.; Kise, Y.; Nozawa, M.; Kuwada, C.; Muramatsu, C.; Katsumata, A.; Fujita, H.; Ariji, E. Performance of deep learning object detection technology in the detection and diagnosis of maxillary sinus lesions on panoramic radiographs. *Dentomaxillofac. Radiol.* **2020**, *50*, 20200171. [CrossRef]
25. Ozbay, S.; Tunc, O. Deep Learning in Analysing Paranasal Sinuses. *Elektron. Ir Elektrotechnika* **2022**, *28*, 65–70. [CrossRef]
26. Bhattacharya, D.; Becker, B.T.; Behrendt, F.; Bengs, M.; Beyersdorff, D.; Eggert, D.; Petersen, E.; Jansen, F.; Petersen, M.; Cheng, B.; et al. Supervised Contrastive Learning to Classify Paranasal Anomalies in the Maxillary Sinus. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13433, pp. 429–438. [CrossRef]
27. Hwang, I.K.; Kang, S.R.; Yang, S.; Kim, J.M.; Kim, J.E.; Huh, K.H.; Lee, S.S.; Heo, M.S.; Yi, W.J.; Kim, T. Il SinusC-Net for automatic classification of surgical plans for maxillary sinus augmentation using a 3D distance-guided network. *Sci. Rep.* **2023**, *13*, 11653. [CrossRef]
28. Xu, H.; Xu, Q.; Cong, F.; Kang, J.; Han, C.; Liu, Z.; Madabhushi, A.; Lu, C. Vision Transformers for Computational Histopathology. *IEEE Rev. Biomed. Eng.* **2024**, *17*, 63–79. [CrossRef]
29. Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; Hong, Q. LViT: Language Meets Vision Transformer in Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2024**, *43*, 96–107. [CrossRef]
30. Smith, P.; Reid, D.B.; Environment, C.; Palo, L.; Alto, P.; Smith, P.L. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *20*, 62–66.
31. Murata, M.; Ariji, Y.; Ohashi, Y.; Kawai, T.; Fukuda, M.; Funakoshi, T.; Kise, Y.; Nozawa, M.; Katsumata, A.; Fujita, H.; et al. Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. *Oral Radiol.* **2019**, *35*, 301–307. [CrossRef]
32. Mori, M.; Ariji, Y.; Katsumata, A.; Kawai, T.; Araki, K.; Kobayashi, K.; Ariji, E. A deep transfer learning approach for the detection and diagnosis of maxillary sinusitis on panoramic radiographs. *Odontology* **2021**, *109*, 941–948. [CrossRef]

33. Kotaki, S.; Nishiguchi, T.; Araragi, M.; Akiyama, H.; Fukuda, M.; Ariji, E.; Ariji, Y. Transfer learning in diagnosis of maxillary sinusitis using panoramic radiography and conventional radiography. *Oral Radiol.* **2023**, *39*, 467–474. [CrossRef]
34. Altun, O.; Özen, D.Ç.; Duman, Ş.B.; Dedeoğlu, N.; Bayrakdar, İ.Ş.; Eşer, G.; Çelik, Ö.; Sümbüllü, M.A.; Syed, A.Z. Automatic maxillary sinus segmentation and pathology classification on cone-beam computed tomographic images using deep learning. *BMC Oral Health* **2024**, *24*, 1208. [CrossRef]
35. Bayrakdar, I.S.; Elfayome, N.S.; Hussien, R.A.; Gulsen, I.T.; Kuran, A.; Gunes, I.; Al-Badr, A.; Celik, O.; Orhan, K. Artificial intelligence system for automatic maxillary sinus segmentation on cone beam computed tomography images. *Dentomaxillofac. Radiol.* **2024**, *53*, 256–266. [CrossRef]
36. Çelebi, A.; Imak, A.; Üzen, H.; Budak, Ü.; Türkoğlu, M.; Hanbay, D.; Şengür, A. Maxillary sinus detection on cone beam computed tomography images using ResNet and Swin Transformer-based UNet. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2024**, *138*, 149–161. [CrossRef]
37. Ministry of Health. Kingdom of Saudi Arabia, Ministry of Health. Available online: https://www.moh.gov.sa/en/Pages/default.aspx (accessed on 8 December 2024).
38. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
39. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 10691–10700.
40. Gao, L.; Zhang, J.; Yang, C.; Zhou, Y. Cas-VSwin transformer: A variant swin transformer for surface-defect detection. *Comput. Ind.* **2022**, *140*, 103689. [CrossRef]
41. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [CrossRef]
42. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.