Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.040251

ARTICLE

SmokerViT: A Transformer-Based Method for Smoker Recognition

Ali Khan[1,4], Somaiya Khan[2], Bilal Hassan[3], Rizwan Khan[1,4] and Zhonglong Zheng[1,4,*]

[1]College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua, 321004, China

[2]School of Electronics Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China

[3]Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, 127788, United Arab Emirates

[4]Key Laboratory of Intelligent Education of Zhejiang Province, Zhejiang Normal University, Jinhua, 321004, China

*Corresponding Author: Zhonglong Zheng. Email: zhonglong@zjnu.edu.cn

Received: 10 March 2023 Accepted: 01 July 2023 Published: 31 October 2023

ABSTRACT

Smoking has an economic and environmental impact on society due to the toxic substances it emits. Convolutional Neural Networks (CNNs) need help describing low-level features and can miss important information. Moreover, accurate smoker detection is vital with minimum false alarms. To answer the issue, the researchers of this paper have turned to a self-attention mechanism inspired by the ViT, which has displayed state-of-the-art performance in the classification task. To effectively enforce the smoking prohibition in non-smoking locations, this work presents a Vision Transformer-inspired model called SmokerViT for detecting smokers. Moreover, this research utilizes a locally curated dataset of 1120 images evenly distributed among the two classes (Smoking and NotSmoking). Further, this research performs augmentations on the smoker detection dataset to have many images with various representations to overcome the dataset size limitation. Unlike convolutional operations used in most existing works, the proposed SmokerViT model employs a self-attention mechanism in the Transformer block, making it suitable for the smoker classification problem. Besides, this work integrates the multi-layer perceptron head block in the SmokerViT model, which contains dense layers with rectified linear activation and linear kernel regularizer with L2 for the recognition task. This work presents an exhaustive analysis to prove the efficiency of the proposed SmokerViT model. The performance of the proposed SmokerViT performance is evaluated and compared with the existing methods, where it achieves an overall classification accuracy of 97.77%, with 98.21% recall and 97.35% precision, outperforming the state-of-the-art deep learning models, including convolutional neural networks (CNNs) and other vision transformer-based models.

KEYWORDS

Smoker recognition; SmokerViT; deep learning; transformer for vision

1 Introduction

The smoking epidemic is one of the world's significant public health threats, killing more than 8 million people yearly, including 1.2 million from passive smoking. In 2020 statistics [1], 22.3% of the

world's population smoke, and 80% of those 1.3 billion smokers worldwide are in low- and medium-income countries. According to a report about different causes of death worldwide, smoking is the second most significant risk factor for death [2]. Smoking monitoring and preventive policies are included as actions that should be implemented in the World Health Organization (WHO) framework convention on smoking control [3]. Therefore, detecting smokers in no-smoking areas is essential for effective surveillance.

Traditional surveillance methods for smoker detection are inefficient and affected by various factors, thus limiting the development of intelligent surveillance [4,5]. Researchers have continuously applied different methods to surveillance systems to answer these problems and benefit from artificial intelligence technology [6,7]. Deep learning is the state-of-the-art (SOTA) artificial intelligence method that has become integral to computer vision [8]. Compared to traditional image processing and machine learning methods, deep learning does not require complex image pre-processing. The Deep Neural Network (DNN), which employs deep learning techniques [9], significantly improves object detection efficiency by automatically learning the features from raw data. Convolutional Neural Networks (CNNs), a subset of DNNs, have been widely utilized to classify and cluster images based on similarity and recognize objects in scenes. CNNs have encouraged the exponential rise in deep learning as it enables significant advancements in many exciting applications, including surveillance [10,11], medical diagnosis [12], self-driving cars [13], etc.

Since the first CNN model AlexNet [14], resulted in faster training time efficiency, new CNN models are proposed with improved accuracy with fewer parameters. Early models, such as the Visual Geometry Group (VGG) [15], had many neurons and parameters, which may result in overfitting and involve enormous computational resources. With the application of residual blocks, the training efficiency of CNN models was improved with some widely used SOTA models such as ResNet [16], Inception [17], and DenseNet [18]. All the previously proposed CNN methods employed in different applications showed that accuracy is critical for applying deep learning in computer vision.

CNN models have become a vital tool in computer vision-based surveillance applications. Convolution layers were previously widely used as the fundamental building block; however, current trends of adding attention processes have prompted researchers to rethink this. In addition to assisting CNNs with long-range dependencies, attention may replace convolutions to provide SOTA performance on computer vision tasks [19]. Recently, researchers have examined using self-attention in vision-based tasks because of its potential for word-dependency learning abilities [20]. Self-attention helps to learn complex relations between neighbours and their further neighbours, which may help with the binary classification problem.

CNNs need help in describing low-level features and can miss important information. Moreover, accurate detection is vital with minimum false alarms. Considering the need for Artificial intelligence (AI) based surveillance mechanism for smoker recognition in no-smoking indoor and outdoor environments, this study focuses on the interpretation of self-attention and multilayer perceptron (MLP) head for a better understanding of the employed deep learning method. To accurately recognize smokers, this work introduces the method SmokerViT for smoker recognition in smart city indoor and outdoor environments where the Transformer component learns the convolution-like features. In SmokerViT, the patch extractor is 16×16 convolution with stride 16. Its output is then multiplied by learnt weights to form q, k, and v embeddings of the self-attention layer. Moreover, the MLP residual block is a linear layer in SmokerViT that raises the embedding dimension by a factor of 4, adds non-linearity then lowers it back to the original. Further linear kernel $L2$ classifier is used for classification. The proposed SmokerViT model uses these threefold attributes, resulting in a more

robust recognition system for Smoking and NotSmoking images. The novelty of this research is to develop a deep learning-based SmokerViT model for Smoker recognition with higher accuracy. The main contributions of this research are:

- This research utilizes the smoker detection dataset, which has 1120 images evenly distributed among the two classes (Smoking and NotSmoking). This research performs augmentations on the dataset to have a considerable number of images with various representations to overcome the dataset size limitation.
- This paper presents a novel end-to-end deep learning model called SmokerViT, which integrates transformer blocks and MLP head with a fully connected layer to learn complex features and linear kernel 2 regularizer for recognizing smokers. Moreover, SmokerViT, due to its discriminative multi-head self-attention, possesses the intrinsic capabilities to classify images irrespective of the backgrounds, image quality, and noisy artefacts.
- This research performs exhaustive analysis to optimize the SmokerViT model to achieve the best performance on the test dataset. It can facilitate future research as a starting point for efficient smoker recognition methods.
- The performance of the proposed SmokerViT model is compared with different deep-learning models on the smoker detection dataset. This work uses several evaluation metrics to assess the performance of the SmokerViT model, where it outperforms the existing state-of-the-art solutions in classification accuracy.

The research paper is organized as follows: Section 2 details the related work associated with this research, Section 3 gives the details of materials and methods adopted for solving the recognition task for the smoker detection problem, and Section 4 offers a detailed performance analysis of the proposed method and comparison with other methods and Section 5 concludes this research.

2 Related Work

There has been some research on various applications of surveillance using computer vision based on different proposed CNN methods. These computer vision applications include human activity recognition, pedestrian detection, traffic monitoring, face recognition, vehicle identification, fire detection, motion detection, medical imaging, etc. Authors in [21] compared state-of-the-art machine learning algorithms for insurance fraud detection. The proposed study's decision tree algorithm performed best for the considered task. Similarly, authors in [22] proposed an improved particle swarm optimization method for data classification. Their proposed method has been tested to optimize the weight of the feed-forward neural network for fifteen datasets. Another research [23] proposed CNN based model for person head detection for counting the crowd in sports videos. Their proposed method solves the multi-scale problem, which is the object detection problem's core issue.

The smoker detection problem is relatively new and less explored, possibly due to the unavailability of open-access image/video datasets. Authors in [24] proposed a deep learning method based on YOLOv3-tiny named Improved YOLOv3-tiny to solve the problem of indoor low-precision smoke alarms on their local dataset. The proposed method combined the advantages of YOLOv3 and YOLOv3-tiny in terms of fewer parameters and higher accuracy for the localization task on their local smoker dataset. The proposed method considered the performance metrics of mAP for the localization task. It showed 85% mAP for improved YOLOv3-tiny compared to YOLOv3-tiny, which was 74%. However, their work limitations are the low mAP and the unavailability of the dataset. Another similar method [25] was proposed, named Eye-Smoker, YOLOv3 based transfer learning method for smoker detection on their local dataset. In the proposed method, the smoker is detected based on the cigarette

and does not consider other kinds, such as e-cigarettes and smoking pipes. Their proposed method considered the localization task for smoker detection with 90% accuracy and around 94% mAP. Their limitations of work are the low accuracy and unavailability of the dataset. These object detection methods promote fast localization capabilities but lack high accuracy.

For the classification problem, false alarms should be kept minimal. A significantly high rate of false alarms in one class and a higher number of accurate classifications in another can lead to higher prediction accuracy; however, it might lack solving the desired problem. In [26], the authors proposed a SmokingNet model based on GoogleNet for smoker detection problems on their local dataset. Their work focused on evaluating the performance of smoking and not-smoking image classification with different performance metrics. In their proposed method, the smoking image characteristics are optimized based on the GoogleNet, and the feature extraction ability is enhanced using kernels of non-square convolution. The proposed method achieved 90% accuracy, 90% precision and recall, and 90% F1 measure. Their work limitations are using very basic GoogleNet as a base model and the unavailability of the dataset. In previously published work [27], the research proposed Inception-ResNet-V2-based transfer learning, where the pre-trained model was used as a backbone network for the smoker detection problem on the local smoker detection dataset. In the proposed method, the Inception-ResNet-V2 model is used, which is trained on the ImageNet dataset, the weights of the pre-trained Inception-ResNet-V2 are frozen, and new fully connected layers are added with ReLU and sigmoid activation functions. The fully connected layers learn the specific features of the task of smoker detection. The proposed method fed the complete image with an input size of 224×224 to the network. The neural network extracted the features based on the previously learned generic features trained on the ImageNet dataset. The proposed solution has a training accuracy of 95.65% and 96.87% testing accuracy with a recall of 97.32% and precision of 96.46%, discriminating the images of the Smoking and NotSmoking classes. However, the proposed work had high accuracies; still, it lacked training the model from scratch and better learn the low-level features.

To solve the parallel processing of words by using self-attention in Recurrent Neural Network (RNN) models, a network called Transformer based on attention mechanism and removes recurrence and convolutions was proposed [28], which accomplished great success in natural language processing (NLP). After its success in NLP, an image classification model, Vision Transformer (ViT) [29], was introduced in computer vision, disrupting the traditional CNN model with its competitive performance on large-scale image datasets. With the development of transformers for computer vision in 2021, there has been some research for computer vision applications using vision-based transformers [30–32]. Transformers have seen much growth in image classification tasks with accuracy similar to if not more than, CNN models. In [33], the authors proposed a multi-instance vision transformer named MITformer for remote sensing scene classification. In their proposed method, the local feature response was highlighted for the remote sensing scenes. Attention-based MLP was inserted at the end of each encoder to enhance these features. Another work [34] proposed a hybrid CNN and ViT method, CTNet, to classify high-resolution remote sensing (HRRS) images. The proposed method has two modules, T-stream (stream for ViT) and C-stream (stream for CNN). In the T-stream, the flattened patches of the image are sent into the pre-trained ViT for semantic features in HRRS images. At the same time, C-stream is used to extract the local features. Ma et al. [35] proposed a homo-heterogeneous transformer learning (HTTL) for remote sensing scene classification. In the proposed HTTL, a patch generation module is used to design homo- and heterogeneous patches. The feature learning module extracts the feature information of global and local areas. A fusion submodule and metric learning-based classification module are used for the scene classification.

In other computer vision applications, authors [36] proposed detecting rain and road surface conditions using vision transformers. In their proposed method, a spatial self-attention network is proposed to analyze the relationship between the detection results of adjacent images as a sequence-to-sequence detection task. Dong et al. [37] proposed ViT based representation learning method for polarimetric synthetic aperture radar (PolSAR) image classification. In the proposed method, the ViT learned the global features of the PolSAR images, which improves the classification efficiency. In [38], the authors proposed a multilabel vision transformer ForestViT for multilabel classification of satellite images of deforestation, which adopts a self-attention mechanism, replacing the convolution operations. Wang et al. [39] proposed a double output vision transformer (DOViT) for air quality classification. The tokens are processed with multilabel self-attention (MSA) to extract features for higher accuracy. Authors [40] proposed Transformer based LPViT for classifying and detecting defects in printed circuit boards (PCBs). The proposed method used labels for better model strategy and mask patch prediction to ensure the relationship of different patch extractions.

However, deep learning-based algorithms were formerly thought of as a black box, and there have been issues with their interpretability for a long time [41–43]. CNNs face problems describing the low-level features outside the actual area of interest [44]. Taking advantage of the context information for feature extraction is not beneficial. This work exploits self-attention blocks and MLP head, which are building blocks of the proposed method for visually interpreting Smoking and NotSmoking images. In the proposed method SmokerViT, which is based on Transformer and MLP head, the smoker recognition problem is considered. It achieves better prediction accuracy without convolutions than the previously proposed CNN methods.

3 Materials and Methods

This section details the proposed SmokerViT and the image dataset used for Smoking and NotSmoking classes for recognition problems. The following subsections explain the methodology of this work.

3.1 Dataset Acquisition

The dataset for this study is the smoker detection dataset published online as open access, which has different images of people smoking and not smoking indoors and outdoors. To the best of our knowledge, there is no other open-access dataset related to the problem; the smoker detection dataset facilitates future work in proposing new methods. The dataset can be accessed from [27].

3.2 Dataset Distribution

The smoker detection problem is considered binary, with two classes named, Smoking and NotSmoking. The NotSmoking class images are labeled 0, while the Smoking images have a class label 1. The smoker detection dataset is balanced and has 1120 images, with 560 images each in the Smoking and NotSmoking classes. This research splits the dataset into training and testing with a ratio of 80:20 with equal distribution from both classes. The training data is further divided into training and validation, with 716 images belonging to training samples and 180 for validation.

3.3 Proposed Method

Smoker detection in no-smoking areas is a difficult task where many factors influence the development of an AI-based surveillance system. The smoker recognition problem is solved by using a smoker detection dataset. Higher detection accuracy needs a large dataset for training, although

applying deep learning models significantly improves the results. If the dataset is too small, the model is at risk of over-fitting, which means it cannot generalize effectively and will result in poor performance on a new dataset. Therefore, to train a small dataset for deep learning, this work performed data augmentations to have multiple training samples to overcome the dataset size limitation. This research implemented various augmentation processes on the training dataset, as given in Table 1. This work performed various augmentations such as resizing, scaling, flipping, shifting, etc., as illustrated in Fig. 1. All the images in the dataset are resized to a uniform resolution of 224 × 224. After that, augmentations are applied, such as vertical and horizontal shift by a factor of 0.2, 50° rotation, zoom by a 0.2 factor, shear transformation, and horizontal flip by 0.2 factor. Some sample augmentation images are depicted in Fig. 1.

Table 1: Data augmentation

Augmentation	Value
Vertical and horizontal shift	0.2
Rotation	50°
Zoom	0.2
Shear transformation	0.2
Horizontal flip	0.2

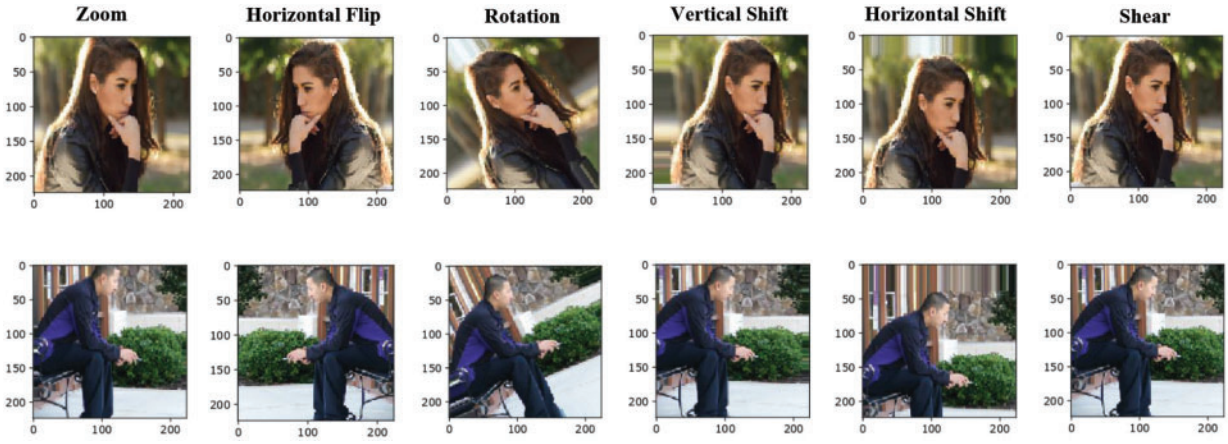


Figure 1: Sample data augmentations

To efficiently execute the recognition task, this research proposes SmokerViT inspired by Vision Transformer [29]. CNNs have been pivotal in solving the problems of computer vision-based applications. In CNN, the pixels of the image are interdependent, and instead of all pixels features being trained on, only extracted features from the image patches using filters are being used for training. However, if complete data of images are used for training, the chances of obtaining the best performance become higher, which is the main work of the Transformer for vision-based applications. In the proposed SmokerViT, the work first converted the image into patches of size 16 × 16. In Transformer, the patches should be of the size that gives the equal rows and columns of the patches in the image. The image size of 224 × 224 and 16 × 16 patch size will give 14 × 14, 196 patches per image. After the conversion into patches, it is passed to the Transformer encoder for processing. After that,

the output is passed into the Multi-Layer Perceptron (MLP) head, which in the proposed SmokerViT consists of flatten, dense layer with ReLU function and a classification layer with kernel regularizer $L2$ to output the prediction depicted in Fig. 2.

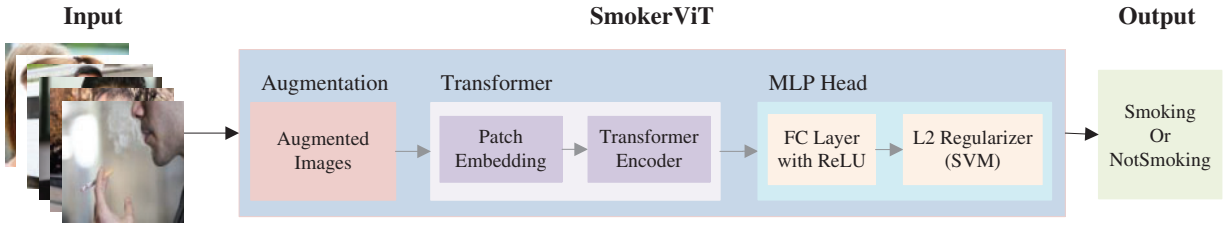


Figure 2: Working mechanism of SmokerViT

In SmokerViT, Vision Transformer takes the series of patches of images as input and predicts the class labels for the input image. The transformer differs from traditional CNNs, which do computations using pixel arrays. The Transformer divides the image into patches of fixed size. Then it inputs these patches into a linear projection of flattened patches embedding layer to produce vectors often known as tokens. These tokens precede a series of tokens. Additionally, the location data is provided by the position embedding. The Transformer encoder will receive these tokens as embedded patches and the location data. The Transformer encoder has the same number of outputs as inputs. The output corresponding to the class is then entered into the MLP head to output the prediction and classification. The architecture of SmokerViT is illustrated in Fig. 3.

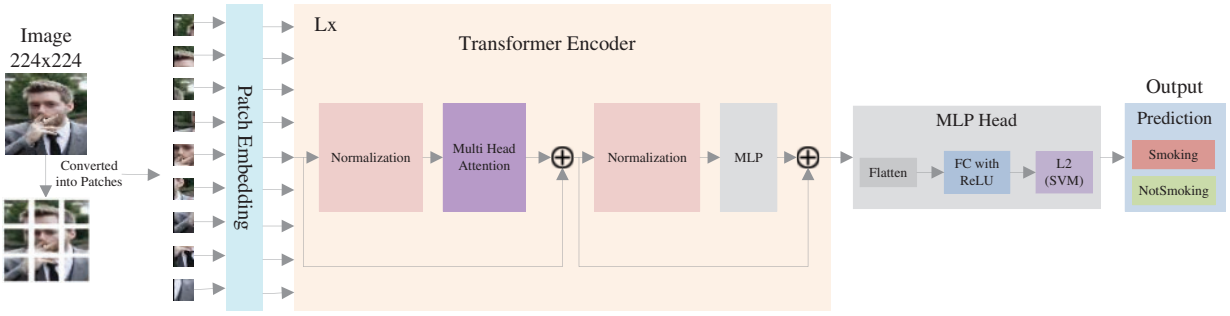


Figure 3: Architecture of SmokerViT

To consider the operation of SmokerViT in detail; first, the input image X with dimension $h \times w \times c$ is divided into several patches of X_p as $n \times (p^2.c)$, where h and w represent the image resolution of input while (p, p) represents the image patch resolution, c denotes the number of channels, and $n = \dfrac{hw}{p^2}$ represents the number of image patches, and this is the input sequence length for the model. These patches are then passed through a linear projection and mapped to the d dimension to get the output referred to as patch embedding. The position embedding $\boldsymbol{E}_{pos}$ is added to the patch embedding $\boldsymbol{E}$ to keep the position information of the input. It is expressed as $\boldsymbol{E}_{pos} \in \mathbb{R}^{(n+1)\times d}$, which joins the [*class*] token $Z_0^0 = X_{class}$. Its form at the output of Transformer encoder Z_L^0 works as image representation Y. The Transformer encoder contains multi-head self-attention (MSA), layer normalization (LN) and MLP block.

MSA: This layer linearly integrates the attention output. The encoder receives a sequence of embedding to process, which undergoes three different linear transformations to output the three

vectors query q, key k, and value v. The attention output for each embedding is calculated by the dot product of these three vectors. Self-attention is calculated independently and repeatedly in parallel. As a result, it is known as multi-head attention. The attention measures how strongly the patches are connected, subsequently assisting in prediction. The MSA is calculated by the equation given below:

$$Attention = softmax\left(\frac{qk^t}{\sqrt{d}}\right)v \tag{1}$$

$$Head_i = Attention(qw_i^q, kw_i^k, vw_i^v) \tag{2}$$

$$MSA = Concat(Head_1, \ldots, Head_i)w^0 \tag{3}$$

where d is the dimension of k, and w_i is the learnable weights.

LN: Layer normalization balances the mean-variance of each input neuron layer, making it converge faster. Layer normalization is added before each block, as it has no prior image dependencies, so it enhances the performance and decreases the execution time.

MLP: The MLP in the Transformer encoder consists of two layers with GeLU.

MLP head: After the Transformer encoder, the output is inserted into the newly added MLP head for the classification of Smoking and NotSmoking images, which consists of flatten layer to flatten the encoder output, dense layers with ReLU activation and linear kernel $L2$ regularizer as a classifier.

3.3.1 Activation Function

The activation function optimizes the processes and learns complex features specific to the task. The proposed method considers ReLU (R) activation function. R is a piecewise linear function that outputs the input directly if it is positive; otherwise outputs zero and is given by:

$$R(X) = \max(0, X) \tag{4}$$

3.3.2 Optimizer

This work considers RMSProp (Root Mean Squared Propagation) optimizer for the proposed SmokerViT method. RMSprop applies the exponential moving average of the squared gradients to adjust the learning rate. RMSprop only accumulates gradients in a specific fix window instead of letting all the gradients accumulate for momentum. The equation for RMSprop is as follows:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{5}$$

where η shows the learning rate, ϵ represents the small term preventing division by zero, $E[g^2]$ is the past squared gradients RMSprop running average, and g_t is the gradient function.

3.3.3 Loss Function

The binary cross entropy loss function is often used for binary classification. It helps to evaluate model accuracy by calculating the prediction probability. Following is the equation for the binary cross entropy loss function:

$$Binary\ Crossentropy\ Loss = z.\log(\hat{z}) + (1 - z).\log(1 - \hat{z}) \tag{6}$$

where z represents the label, i.e., 1 denotes the Smoking class and 0 denotes NotSmoking class, and $\hat{z}$ is the predicted probability of z.

3.3.4 Linear Kernel L2 as a Classifier

This work used the linear kernel $L2$ algorithm because it can help solve problems with multi-collinearity (highly correlated independent variables) by limiting the coefficient and maintaining all the variables. Linear kernel, basic kernel, is the best in case of many features and given by $f\left(X, X_j\right) = \sum X \cdot X_j$ where X, X_j is the data to classify. Linear kernel $L2$ predicts based on the mean of data to avoid overfitting, unlike $L1$, which takes the median of data for estimation. $L2$ adds the penalty to the cost function as the squared value of the weights and learns complex patterns. $L2$ is computationally efficient, and predictions are more accurate when the output is a function of all input variables. $L2$ regularization is calculated by:

$$L2\ regularizer = \lambda \sum_{i=0}^{n} w_i^2 \tag{7}$$

where w_i is the weight and λ represents the regularization parameter. If λ is 0, this acts as Ordinary Least Square (OLS), where it will make the weight coefficient 0 and result in underfitting, while if λ is very large, it will increase the weight and result in underfitting.

Pseudocode:

// Step 1: Data Preparation

dataset = preprocess_images(labeled_images) // [$N \times H \times W \times C$] array of preprocessed images

training_set, validation_set, test_set = split_dataset(dataset) // [N_{train}, H, W, C], [N_{val}, H, W, C], [N_{test}, H, W, C] arrays

// Step 2: Model Architecture

model = create_SmokerViT(num_transformer_blocks, embedding_size, num_attention_heads) // a SmokerViT model

// Step 3: Training

initialize_weights(model) // initialize model weights randomly

for epoch in 1 to num_epochs:

  for batch in training_set:

    loss = calculate_loss(model, batch) // calculate binary cross-entropy loss between model predictions and ground truth labels

    update_weights(model, loss) // update model weights using RMSprop

  validation_accuracy = evaluate(model, validation_set) // calculate validation accuracy

  if validation_accuracy does not improve for num_epochs_to_stop:

    break // Early stopping if validation accuracy plateaus

  end

  end

end

// Step 4: Hyperparameter Tuning

hyperparameters = {learning_rate: [0.01, 0.0001], batch_size: [16, 32, 64], num_transformer_blocks: [6, 12]} // candidate hyperparameters

best_hyperparameters = grid_search(model, hyperparameters, training_set, validation_set) // find best hyperparameters using grid search

// Step 5: Evaluation

test_accuracy = evaluate(model, test_set) // calculate test accuracy

metrics = calculate_metrics(model, test_set) // calculate other evaluation metrics (e.g., precision, recall, F1 score)

---

4 Performance Evaluation

The performance of the proposed SmokerViT for smoker recognition is evaluated and compared with other methods on the smoker detection dataset. The system configurations for simulation are i7-11800H, 16 GB DDR4, NVIDIA RTX3060 6 GB, and the simulation setup is Anaconda Python 3.8 with Tensorflow 2.6 and Keras 2.3 libraries. The proposed SmokerViT is tested with various hyper-parameters values for the best results. Table 2 depicts the hyper-parameters for the simulations.

Table 2: Simulation parameters

Parameters	Value
Input size	224×224
Patch size	16×16
Maximum epochs	50
Batch size	32
Learning rate	1e-3
Loss function	Binary cross entropy
Optimizer	RMSprop

4.1 Evaluation Metrics

This section presents the evaluation metrics for analyzing the performance of the proposed SmokerViT method. This work evaluated the methods on the following metrics:

$$Prediction\ Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{8}$$

$$Precision/Positive\ predictive\ value\ (PP_v) = \frac{T_p}{T_p + F_p} \tag{9}$$

$$Sensitivity/Recall/True\ positive\ rate(TP_r) = \frac{T_p}{T_p + F_n} \tag{10}$$

$$Specificity/Selectivity/True\ negative\ rate(TN_r) = \frac{T_n}{T_n + F_p} \tag{11}$$

$$False\ positive\ rate(FP_r)/Fall\ out = 1 - Specificity \tag{12}$$

$$False\ negative\ rate(FN_r)/miss\ rate = 1 - Sensitivity \tag{13}$$

$$FD_r = 1 - PP_v \tag{14}$$

$$F1 = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right) \tag{15}$$

$$E_r = \frac{F_p + F_n}{T_p + T_n + F_p + F_n} \tag{16}$$

The T_n is the true negative, accurately classified as NotSmoking images, while T_p is the true positive, accurately classified as Smoking images by the proposed solution. F_n is the false negative where the Smoking image is categorized as NotSmoking, and F_p is the false positive where NotSmoking

images are labelled as Smoking. *Precision* is the percentage of correctly positive outcomes to those the classifier predicted correctly, whereas the ratio of correctly positive results to all the relevant samples that should be positive is known as the *Recall* or *Sensitivity* of the proposed method. The ratio of correct negative predictions to the results that the classifier predicts as negative is known as *Specificity* or true negative rate. FD_r is the total number of false positive classifications to the total positive classifications. The F1 score is the harmonic mean of precision and recall, which shows how the classifier predicts correctly. FP_r is the ratio of negatives falsely categorized as positives and the total number of actual negatives, while FN_r is the ratio of positives being falsely classified as negatives and the total number of actual positives. E_r is the error rate, the ratio of all the incorrect predictions to the total number of test samples.

4.2 Attention Maps of Learned Features

This subsection interprets the visualization of the proposed method to understand the smoker recognition mechanism better. This research visualized the attention maps of some sample images predicted for the smoker recognition tasks from the self-attention block, illustrated in Fig. 4. Self-attention is the main reason the Transformer integrates data across the complete image, including in the base layers. The attention maps show how well the method utilizes this capacity for the considered task. Some attention heads already focus on the desired representations on most images in the initial layers, demonstrating that the proposed method employs the capability to integrate information globally. The attention to the desired features increases with the model depth, and it becomes clearer what features the model pays attention to for the desired task. The original input images are converted into pseudo-colour images to highlight the attention mask applied to the input image. It can be seen from the sample of images considered for both the Smoking and NotSmoking classes that the brighter part represents the attention mapping of the proposed method. Globally this research discovers that the method pays attention to image areas that are significant for classification in terms of semantics. From the maps, it can be noted that the attention of the method is on the cigarettes and their smoke for the detection of smokers. Similarly, the absence of cigarettes and smoke around the person's mouth or hand is predicted to be a NotSmoking image.

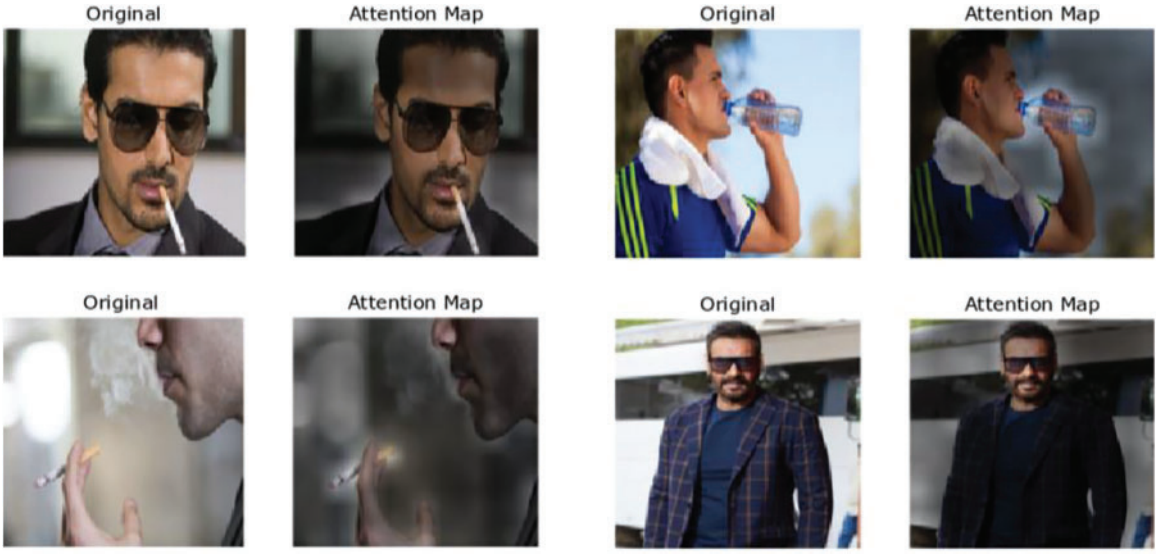


Figure 4: (Continued)

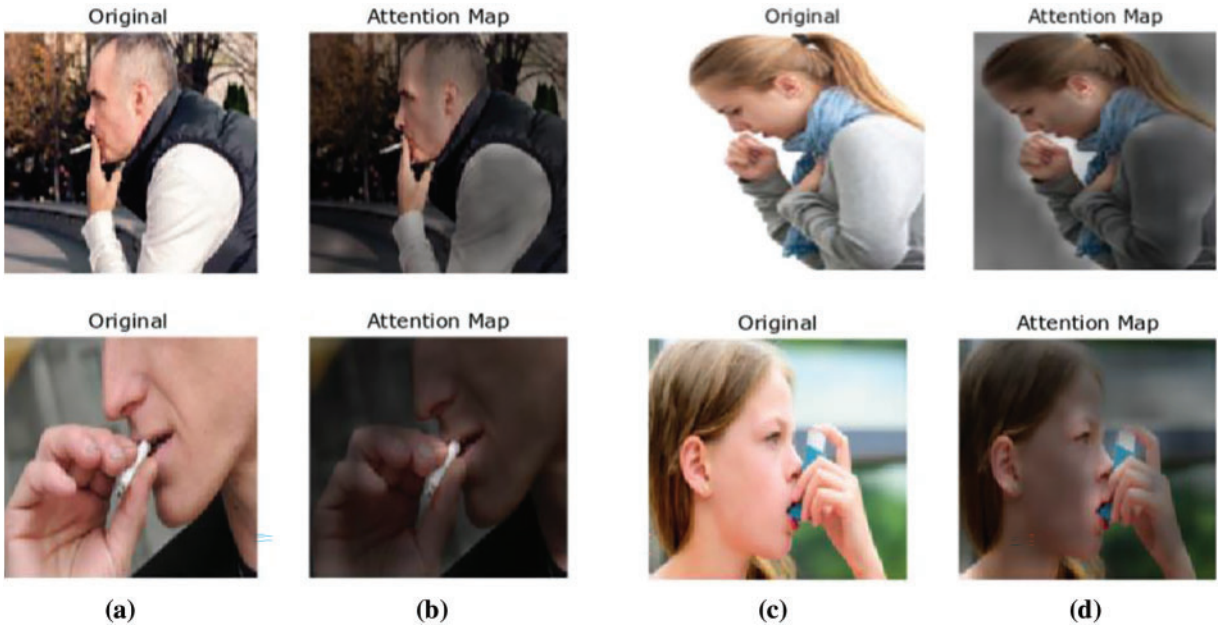


Figure 4: Sample images of smoking and NotSmoking (a, d) original and (b, c) attention maps

4.3 Performance Analysis of SmokerViT

This section presents the performance analysis of the proposed SmokerViT, a method based on the Transformer and MLP head. This work was analyzed by using different regularizers as a classifier to prove the effectiveness of linear kernel regularizer ($L2$) over Gaussian kernels. Moreover, this work proves the efficiency of the proposed method with and without the proposed MLP head and with and without data augmentation with simulation settings presented in Table 2.

Table 3 presents the performance of SmokerViT in terms of prediction accuracy using different kernels. The result shows that using linear kernel $L2$ SmokerViT obtained the best result with 97.77% overall prediction accuracy, while the Gaussian kernel displayed overall prediction accuracy of 94.64%. It can be noted that linear kernel $L2$ showed the best accuracy for the Smoking and NotSmoking classes with 98.21% and 97.32%, respectively, whereas the Gaussian kernel showed 93.75% Smoking and 95.54% accuracy for NotSmoking class.

Table 3: Performance of SmokerViT using different kernels

Class	Linear kernel $L2$	Gaussian kernel
Smoking	98.21%	93.75%
NotSmoking	97.32%	95.54%
Overall	97.77%	94.64%

This work considered the best result obtained on linear kernel $L2$ for the SmokerViT. This research performed further analysis by removing the MLP head and replaced with a sigmoid as a classifier with a single output to demonstrate the efficiency of the proposed model. Table 4 presents the performance of the proposed method with and without MLP head and augmentation and without

augmentation block. It can be observed that the proposed model with MLP head and augmentation has the best results, whereas without augmentation performed worst for both the classes and overall prediction accuracy. The performance of SmokerVit with the proposed MLP head is improved using augmentation to 97.77% from 95.54% without the MLP head. This is because the augmentation significantly increased the dataset size with various representations, which helped predict unseen Smoking and NotSmoking images in the test dataset. While without augmentation, and proposed MLP head has significantly low performance with 88.39% overall prediction accuracy.

Table 4: Performance of SmokerViT with and without augmentation and MLP head

Class	With proposed MLP		Without proposed MLP	
	Without augmentation	With augmentation	Without augmentation	With augmentation
Smoking	92.85%	98.21%	89.29%	94.64%
NotSmoking	91.07%	97.32%	87.50%	96.43%
Overall	91.96%	97.77%	88.39%	95.54%

Moreover, this work performed simulations to show the effectiveness of using a Transformer as the main network over other SOTA models with an $L2$ kernel classifier. This work used ResNet, Inception-ResNet-V2 model to show the efficiency of using Transformer over these models. Table 5 shows that the best results are achieved using Transformer with 97.77% accuracy, while the second best results are achieved using Inception-ResNet-V2 with 96.43% accuracy, followed by InceptionV3 with 87.05% accuracy. ResNet performed worse with 85.71% accuracy. It can be observed from the table that SmokerViT has more parameters than the other models; however, the focus of this study is the higher accuracy.

Table 5: Performance comparison of using different models for feature extraction

Model	Params (mln)	Training time (hr)	Accuracy
Inception-ResNet-V2+L2 kernel	73.99	1.83	96.43%
ResNet152V2+L2 kernel	60.41	4.27	85.71%
InceptionV3+L2 kernel	23.92	1.52	87.05%
SmokerViT	86.19	3.66	97.77%

The time complexity of the proposed SmokerViT model can be expressed as $O(N^2L)$, where N represents the number of patches in the input image, and L represents the number of self-attention layers in the transformer.

The $O(N^2L)$ notation arises from each self-attention layer having a quadratic complexity of $O(N^2)$, as it involves computing pairwise dot products between all pairs of patches. Since the SmokerViT model has L self-attention layers, the total time complexity is $O(N^2L)$.

It is worth noting that the SmokerViT model also has additional computational costs associated with the feedforward network and positional embeddings, but these are typically negligible compared to the self-attention computation.

The loss performance curves of the proposed SmokerViT in terms of training loss and validation loss are depicted in Fig. 5. The training phase of the proposed SmokerViT is carried out through 50 epochs. From the result, it can be noted that the training loss started at 76.24% and achieved less than 19.56% loss at the 10th epoch. After 10 epochs, the loss curve remained steady till the 50 epochs, with a final loss of 8.28%. Similarly, the validation loss started at 30.82% and reached 9.62% at the 10th epoch. After that, the loss curve remained steady till 50 epochs, with a final loss of 5.39%.

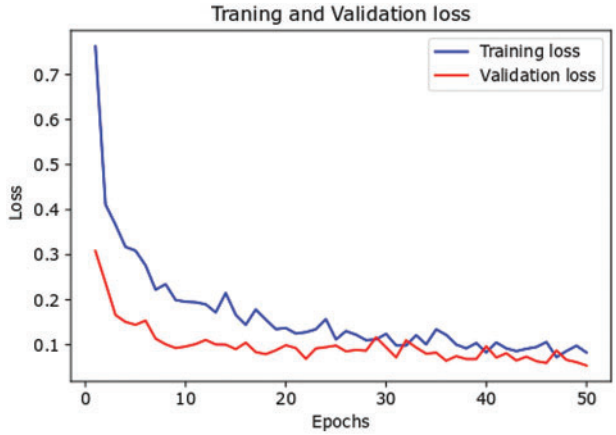


Figure 5: Loss performance of the proposed SmokerViT

The confusion matrix depicts the predictive analysis of Smoking and NotSmoking image classification. It can be seen from the confusion matrix in Fig. 6 that the proposed SmokerViT displayed a prediction accuracy of 97.77% and 2.23% error rate with 109 T_n and 110 T_p with 3 F_p and 2 F_n, respectively. Table 6 shows the performance of the proposed SmokerViT on individual classes. The proposed method achieved 98.21% prediction accuracy, 97.35% precision, 98.21% recall, and 97.78% F1 score for the Smoking class. Whereas for the NotSmoking class, the proposed method displayed 97.32% prediction accuracy, 98.20% precision, 97.32% recall, and 97.76% F1. The proposed SmokerViT displayed the overall performance with 97.77% prediction accuracy, 98.21% recall, 97.35% precision, and 97.78% F1 measure for classifying Smoking and NotSmoking images of the smoker detection dataset.

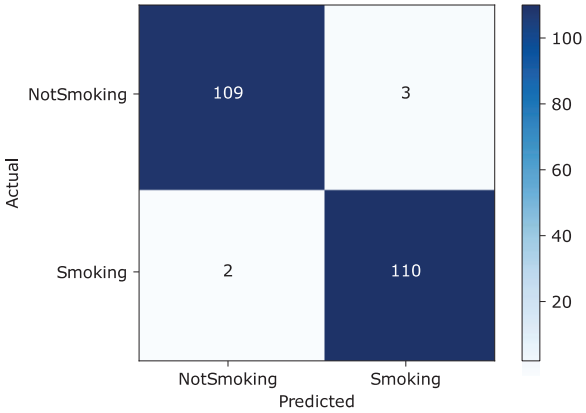
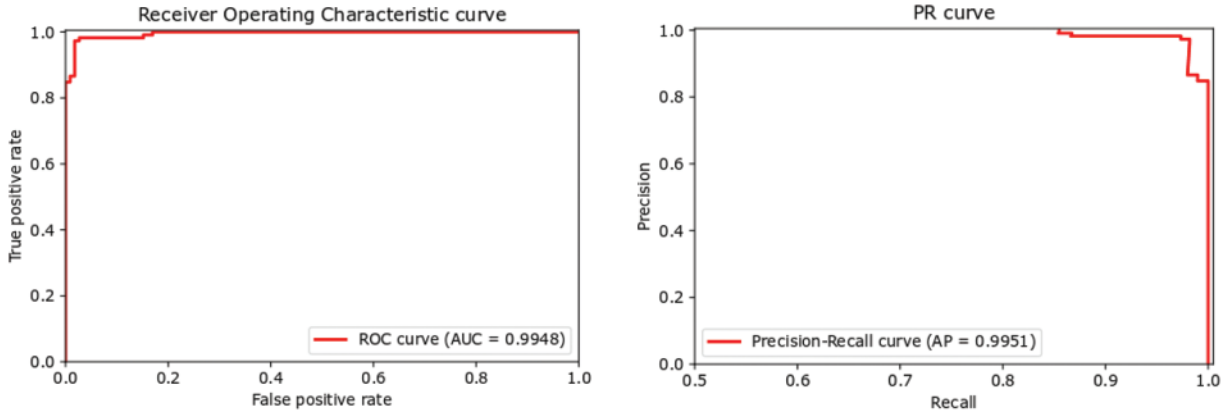


Figure 6: Confusion matrix of SmokerViT

Table 6: Performance of SmokerViT

Class	*Accuracy*	*Precision*	*Recall*	*F*1
Smoking	98.21%	97.35%	98.21%	97.78%
NotSmoking	97.32%	98.20%	97.32%	97.76%
Overall	97.77%	97.35%	98.21%	97.78%

Receiver Operating Characteristic (ROC) curve, shown in Fig. 7a, is another graphical representation for assessing the performance that shows the proposed method's ability to predict classification with varying prediction thresholds. The ROC curve is plotted by considering Recall (TP_r) on the y-axis against FP_r on the x-axis. The Area under the Curve (AUC) depicts how well the method differentiates between the classes. The AUC of 0.9948 by the SmokerViT means that it has a 99.48% chance of accurately classifying the Smoking and NotSmoking classes. This work also analyzed the proposed SmokerViT based on the Precision-Recall (PR) curve, which depicts how well it performed for classifying Smoking images because, unlike ROC, the PR curve does not consider T_n for performance evaluation. PR curve also depicts whenever the class distribution has variation, unlike ROC, which shows no change. It can be noted from Fig. 7b that the curve is near the top right corner, showing that the SmokerViT performed well in classifying the Smoking class. SmokerViT achieved 99.51% average precision (AP) for the Smoking and NotSmoking classification.



(a) **(b)**

Figure 7: (a) Receiver operating characteristic curve with AUC and (b) precision-recall curve with AP

For the smoker recognition problem, false negatives should be minimal. From Fig. 8b, it can be seen that the false negative occurred when the background of the image is similar to the person in the image. In computer vision, spatial resolution is crucial, which has led to the inaccurate classification of the Smoking image as NotSmoking. Better-quality images let the model generalize more accurately. In addition, the neural network had trouble distinguishing between cigarette and the background pixels in the images when the background was blurry. The lack of a considerable number of varying images in the training set might also account for the false alarms. Another possible reason is that some photos in the test set were new to the model and lacked representation of comparable images in the training set. The model performed poorly in generalizing the novel scenarios.

Figure 8: (a) False positives and (b) false negatives

Subsequently, a similar problem was seen with false positives. Smoker recognition relies heavily on the accuracy and practicality of the classifier, both of which are affected by the number of false positives. Images of false positives are shown in Fig. 8a. The scarcity of diversity in the training set and the variety of datasets may result in the incorrect classification of some NotSmoking images as Smoking. It can also be observed image with the background as a cloud was misclassified as Smoking; moreover, a similar hand gesture to the smoking was also labeled as Smoking.

4.4 Comparative Study with Other Methods

For validating the effectiveness of the proposed SmokerViT, this research compared the performance with other methods, both CNN and Transformer based models such as EfficientNetV2 [45], ResNest [46], MobileNetV3 [47], ResNetD [48], ViT [29], Levit [49], Davit [50] and Coatlite [51] on the smoker detection dataset. The hyperparameters are listed in Table 7. Table 8 presents the comparative analysis of these methods. It can be observed that SmokerViT displayed superiority over other considered methods in classifying the Smoking and NotSmoking classes. All the methods considered for comparison were used as pre-trained models using transfer learning and added the classification layer with a sigmoid activation function. After SmokerViT, ViT performed better among all the other considered methods for classification tasks on the local dataset for the smoker detection problem, as explained in Table 6. ViT achieved 96.43% accuracy, 96.43% sensitivity, and 96.43% specificity, followed by Levit with 94.64% accuracy, Coatlite with 91.07% accuracy, Davit with 90.18%,

ResNetD with 88.40% accuracy, ResNest with 86.61% accuracy, EfficientV2 with 85.27% and at last MobileNetV3 with 82.14% accuracy. It can be observed that ResNetD and Davit outperformed ViT in terms of sensitivity at 97.32% compared to ViT at 96.43%. However, MobileNetV3 performed worse in terms of specificity than other models due to the significantly large number of false positives and considerably lower number of false negatives. The proposed method performed best among other considered methods in all evaluation metrics. MobileNetV3 has the lowest accuracy of 82.14% compared to other CNN methods for analyzing the unique smoker classification problem. Fig. 9 shows the performance comparison of all the methods on individual classes. SmokerViT achieves the best results on both classes, followed by ResNetD and Davit for Smoking class with 97.32% accuracy while ViT for NotSmoking class with 96.63% accuracy.

Table 7: Hyperparameters of the models

Parameters	EfficientNetV2	ResNest	MobileNetV3	ResNetD	Levit	Davit	Coatlite	ViT
Structure	CNN	CNN	CNN	CNN	CNN+Transformer	Transformer	Transformer	Transformer
Input image	224 × 224	224 × 224	224 × 224	224 × 224	224 × 224	224 × 224	224 × 224	224 × 224
Patch size	–	–	–	–	16 × 16	16 × 16	–	16 × 16
Epochs	50	50	50	50	50	50	50	50
Batch size	32	32	32	32	32	32	32	32
Learning rate	1e-4	1e-4	1e-3	1e-4	1e-3	1e-3	1e-3	1e-3
Optimizer	Adam	Adam	RMSprop	Adam	RMSprop	RMSprop	RMSprop	RMSprop

Table 8: Comparative analysis of SmokerViT with other methods on smoker detection dataset

Method	*Predictio Accuracy*	*Sensitivity*	*Specificity*	FD_r	FP_r	FN_r
EfficientNetV2 [45]	85.27%	91.07%	79.46%	18.4%	20.54%	8.93%
ResNest [46]	86.61%	90.18%	83.04%	15.83%	16.96%	9.82%
MobileNetV3 [47]	82.14%	92.86%	71.43%	23.53%	28.57%	7.14%
ResNetD [48]	88.40%	97.32%	80.36%	16.79%	19.64%	2.68%
Levit [49]	94.64%	93.75%	95.54%	4.55%	4.46%	6.25%
Davit [50]	90.63%	97.32%	83.04%	14.84%	16.96%	2.68%
Coatlite [51]	91.10%	93.75%	88.39%	11.02%	11.61%	6.25%
ViT [29]	96.43%	96.43%	96.43%	3.57%	3.57%	3.57%
SmokerViT (ours)	**97.77%**	**98.21%**	**97.32%**	**2.65%**	**2.68%**	**1.79%**

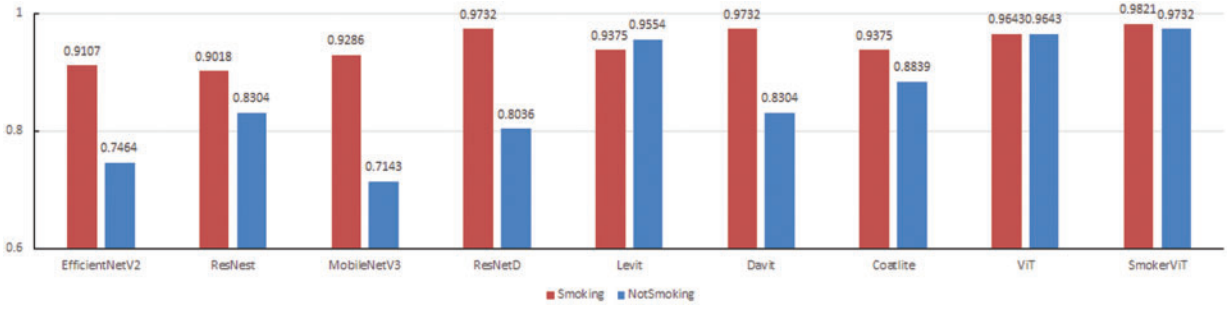


Figure 9: Performance comparison in terms of classes of SmokerViT and other methods

4.5 Comparative Study with Previous Work

This work compares the performance of the proposed SmokerViT with our previous work [27]. Table 9 shows SmokerViT has a better overall prediction accuracy of 97.77% compared to 96.87% by our previously proposed method InceptionResNetV2. It can be noted that SmokerViT has improved the performance of smoker recognition for both classes, where it displayed 98.21% and 97.32% accuracies in discriminating the Smoking and NotSmoking images, respectively. InceptionResNetV2 achieved 97.32% and 96.43% accuracies for the Smoking and NotSmoking classes. Fig. 10 compares the two methods in terms of precision, recall, and F1 measure. It can be observed that SmokerVit performed better in all the considered performance metrics, which is because the self-attention mechanism focuses on the entire image, unlike convolutions in CNN that focus on the interpretation of the high-level features rather than low-level features in classifying Smoking and NotSmoking images.

Table 9: Comparative analysis of SmokerViT with previous work on smoker detection dataset

Class	InceptionResNetV2 [27]	SmokerVit
Smoking	97.32%	98.21%
NotSmoking	96.43%	97.32%
Overall	96.87%	97.77%

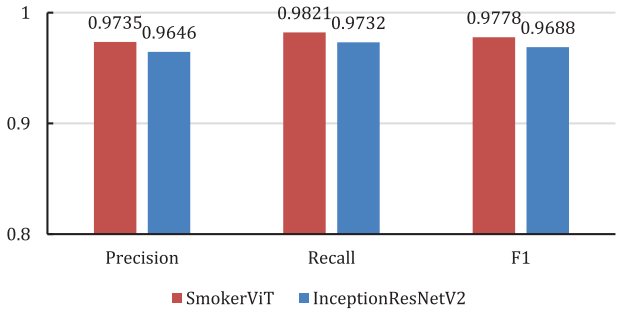


Figure 10: Performance of SmokerViT and InceptionResNetV2

In this research work, SmokerViT displayed self-attention capability and MLP Head to recognize Smoking and NotSmoking images. The results show that better performance is achieved by SmokerViT compared to the SOTA models, implying that the self-attention mechanism and MLP Head architecture may be more suitable than CNN for the Smoker recognition problem. In contrast to transformers, which can compute the attention of any patch, regardless of its distance, a CNN needs to perform additional convolutions to increase the receptive field to determine the relationship between any neighboring pixels, resulting in difficulty in possessing the ability to perform long-range computation. In SmokerViT, the patch embedding component is used to learn convolution-like features, whereas self-attention is used to learn important features and ignore the noisy ones. Results show that the SmokerViT performed better than CNN and Transformer based models, validating the superiority of using both the self-attention mechanism and MLP Head.

While looking at the results, it can be observed that the SmokerViT has performed better in both the Smoking and NotSMoking classes. However, CNNs models were better at predicting Smoking images while poorly classifying NotSmoking class compared to Transformer based models, which

performed well in classifying both the Smoking and NotSmoking classification. While SmokerViT was equally good in classifying both classes indicating that SmokerViT is more robust than using CNN or transformer-based models in dealing with balanced datasets.

Many researchers used CNN for the smoker detection problem, but there is not much work on this problem. This is the first time using transformers and MLP Head with Linear kernel $L2$ classifier for the smoker recognition task. Additionally, previous research results were compared with this proposed work, as depicted in Table 9. It can be observed that SmokerViT outperformed the Inception-ResNet-V2 model for Smoking and NotSmoking image classification using the same dataset. Moreover, the high accuracy displayed by the proposed method can help an AI-based smoker detection system and save time and human resources simultaneously. This research can benefit researchers to improve further the methodology for image segmentation to detect cigarette smoker detection problems.

5 Conclusion

This research proposed a transformer-based smoker recognition method. For effective surveillance of the no-smoking areas, this research proposed SmokerViT based on the self-attention mechanism instead of CNN. The dataset for this work has two classes with 560 images each for the Smoking and Notsmoking classes. Further, this work performed augmentations on the smoker detection dataset to have many images with various representations to overcome the dataset size limitation. The proposed SmokerViT is inspired by Vision Transformer and adding our own MLP head block, which has a dense layer with ReLU activation function and linear kernel $L2$ regularizer as a classifier. SmokerViT extracted features through long-range dependency compared to CNN models, which took advantage of useful global information. Ablations were performed on the proposed SmokerViT to prove the efficiency of the MLP head block and data augmentation. The SmokerViT performance was evaluated and compared with the previously proposed CNN model for the smoker detection problem and other Transformer and CNN-based methods. The SmokerViT achieved a 0.93% higher accuracy of 97.77%, with 0.92% better recall of 98.21% and 0.92% better precision of 97.35% compared to the previous proposed Inception-ResNet-V2 based transfer learning method. Moreover, the results showed that SmokerViT achieved competitive performance compared to other models with considerably higher values of the evaluation metrics.

For future works, several issues that were not addressed in this research need to be observed. The effect of dataset size on the training performance of the method and data augmentation by various complex models for ensuring further variances of the image representations can be considered for future study. Moreover, the weak point of this research is the higher number of parameters and high execution time. In future work, this point would be considered for designing the model, which is lightweight and, at the same time, yield higher accuracy. Moreover, the hybrid method of convolutions and transformer might help the smoker recognition system to perform better, considering the shortcomings of the proposed method.

Acknowledgement: Authors are thankful to the reviewers for their valuable comments.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: A.K.; data collection: A.K.; analysis and interpretation of results: A.K., S.K., B.H. and

R.K.; draft manuscript preparation: A.K., S.K., B.H. and R.K.; writing—review and editing, A.K., S.K., B.H., R.K. and Z.Z.; funding acquisition, A.K. and Z.Z. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset considered in this study can be accessed from https://data.mendeley.com/datasets/j45dj8bgfc/1.

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

References

[1] Report on Tobacco, World Health Organization, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tobacco

[2] Hannah Ritchie and Max Roser, "Smoking," Our World in Data, 2013. [Online]. Available: https://ourworldindata.org/smoking

[3] WHO Framework Convention on Tobacco Control, 2003. [Online]. Available: https://fctc.who.int/who-fctc/overview

[4] N. G. La Vigne, S. S. Lowry, J. A. Markman and A. M. Dwyer, "Evaluating the use of public surveillance cameras for crime control and prevention: A summary," Urban Institute, pp. 1–4, 2011. https://www.urban.org/research/publication/evaluating-use-public-surveillance-cameras-crime-control-and-prevention

[5] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736–753, 2018.

[6] B. R. Raiff, C. Karatas, E. A. McClure, D. Pompili and T. A. Walls, "Laboratory validation of inertial body sensors to detect cigarette smoking arm movements," *Electronics*, vol. 3, pp. 87–110, 2014.

[7] V. Y. Senyurek, M. H. Imtiaz, P. Belsare, S. Tiffany and E. Sazonov, "Smoking detection based on regularity analysis of hand to mouth gestures," *Biomedical Signal Processing and Control*, vol. 51, pp. 106–112, 2019.

[8] A. I. Khan and S. Al-Habsi, "Machine learning in computer vision," *Procedia Computer Science*, vol. 167, pp. 1444–1451, 2020.

[9] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[10] A. Khan, B. Hassan, S. Khan, R. Ahmed and A. Adnan, "DeepFire: A novel dataset and deep transfer learning benchmark for forest fire detection," *Mobile Information System*, vol. 2022, pp. 5358359, 2022.

[11] S. Khan and A. Khan, "FFireNet: Deep learning based forest fire classification and detection in smart cities," *Symmetry*, vol. 14, no. 10, pp. 2155, 2022.

[12] A. S. Panayides, A. Amini, N. D. Filipovic, A. Shama, S. A. Tsaftaris *et al.,* "AI in medical imaging informatics: Current challenges and future directions," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[13] J. Ni, K. Shen, Y. Chen, W. Cao and S. X. Yang, "An improved deep network-based scene classification method for self-driving cars," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[14] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2017.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[16] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1–9, 2015.

[18] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2261–2269, 2017.
[19] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7794–7803, 2018.
[20] J. B. Cordonnier, A. Loukas and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *Proc. of the Int. Conf. on Learning Representation (ICLR)*, Virtual, pp. 1–18, 2020.
[21] L. Rukhsar, W. H. Bangyal, K. Nisar and S. Nisar, "Prediction of insurance fraud detection using machine learning algorithms," *Mehran University Research Journal of Engineering and Technology*, vol. 41, no. 1, pp. 33–40, 2022.
[22] W. H. Bangyal, K. Nisar, T. R. Soomro, A. A. Ag Ibrahim, A. G. Mallah *et al.,* "An improved particle swarm optimization algorithm for data classification," *Applied Sciences*, vol. 13, no. 1, pp. 283, 2023.
[23] S. D. Khan, H. Ullah, M. Ullah, N. Conci, F. A. Cheikh *et al.,* "Person head detection based deep model for people counting in sports videos," in *Proc. of the IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, pp. 1–8, 2019.
[24] R. T. Zhao, M. Y. Wang, Z. L. Zhai, P. Li and Q. Y. Zeng, "Indoor smoking behavior detection based on YOLOv3-tiny," in *Proc. of the Chinese Automation Cong. (CAC)*, Hangzhou, China, pp. 3477–3481, 2019.
[25] J. R. Macalisang, N. E. Merencilla, M. A. D. Ligayo, M. P. Melegrito and R. R. Tejada, "Eye-smoker: A machine vision-based nose inference system of cigarette smoking detection using convolutional neural network," in *Proc. of the IEEE Int. Conf. on Engineering Technologies and Applied Sciences (ICETAS)*, Kuala Lumpur, Malaysia, pp. 1–5, 2020.
[26] D. Zhang, C. Jiao and S. Wang, "Smoking image detection based on convolutional neural networks," in *Proc. of the IEEE Int. Conf. on Computer and Communications (ICCC)*, Chengdu, China, pp. 1509–1515, 2018.
[27] A. Khan, S. Khan, B. Hassan and Z. Zheng, "CNN-based smoker classification and detection in smart city application," *Sensors*, vol. 22, no. 3, pp. 892, 2022.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Proc. of Conf. on Neural Information Processing Systems (NIPS)*, Long Beach, CA, US, pp. 6000–6010, 2017.
[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2021.
[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles *et al.,* "Training data-efficient image transformers & distillation through attention," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, Virtual, pp. 10347–10357, 2021.
[31] Z. Zhang, W. Song and Q. Li, "Dual-aspect self-attention based on transformer for remaining useful life prediction," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.
[32] Q. Luo, J. Su, C. Yang, W. Gui, O. Silvén *et al.,* "CAT-EDNet: Cross-attention transformer-based encoder-decoder network for salient defect detection of strip steel surface," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.
[33] Z. Sha and J. Li, "MITformer: A multiinstance vision transformer for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[34] P. Deng, K. Xu and H. Huang, "When CNNs meet vision transformer: A joint framework for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[35] J. Ma, M. Li, X. Tang, X. Zhang, F. Liu *et al.,* "Homo-heterogenous transformer learning framework for RS scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2223–2239, 2022.
[36] A. Abdelraouf, M. Abdel-Aty and Y. Wu, "Using vision transformers for spatial-context-aware rain and road surface condition detection on freeways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 18546–18556, 2022.

[37] H. Dong, L. Zhang and B. Zou, "Exploring vision transformers for polarimetric SAR image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[38] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 7, pp. 3299–3307, 2023.
[39] Z. Wang, Y. Yang and S. Yue, "Air quality classification and measurement based on double output vision transformer," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 20975–20984, 2022.
[40] K. An and Y. Zhang, "LPViT: A transformer based model for PCB image classification and defect detection," *IEEE Access*, vol. 10, pp. 42542–42553, 2022.
[41] D. Castelvecchi, "Can we open the black box of AI?" *Nature*, vol. 538, pp. 20–23, 2016.
[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh *et al.,* "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 618–626, 2017.
[43] S. Serrano and N. A. Smith, "Is attention interpretable?" arXiv:1906.03731, 2019.
[44] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. of Neural Information Processing Systems*, Virtual, vol. 34, 2021.
[45] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, Virtual, pp. 10096–10106, 2021.
[46] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin *et al.,* "ResNeSt: Split-attention networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 2736–2746, 2022.
[47] A. Howard, M. Sandler, G. Chu, B. Chen, W. Wang *et al.,* "Searching for MobileNetV3," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, South Korea, pp. 1314–1324, 2019.
[48] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie *et al.,* "Bag of tricks for image classification with convolutional neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 558–567, 2019.
[49] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin *et al.,* "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Virtual, pp. 12239–12249, 2021.
[50] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang *et al.,* "DaViT: Dual attention vision transformers," arXiv:2204.03645, 2022.
[51] W. Xu, Y. Xu, T. Chang and Z. Tu, "Co-scale conv-attentional image transformers," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Virtual, pp. 9961–9970, 2021.