Smoker Detection using Vision Models

Project Aim: To detect smokers in public places.

Methodology:

1. *Data Collection:* The dataset contains 1120 images in total, that is divided equally in two classes, in which 560 images belong to smokers and remaining 560 images belong to non-smokers. All the images in the dataset are resized to a resolution of 250*250. Dataset: *Khan, Ali (2022), "Smoker Detection Dataset", Mendeley Data, V1, doi: 10.17632/j45dj8bgfc.1*
2. *Data Preparation:* Image Resizing, Normalization, Label Encoding, and Data Augmentation
3. *Modeling:* VGG16, ResNet-50, MobileNetV2, EfficientNetV2 and Vision Transformer (ViT).
4. *Model Comparison:* Comprehensive analysis of model performance and trade-offs.

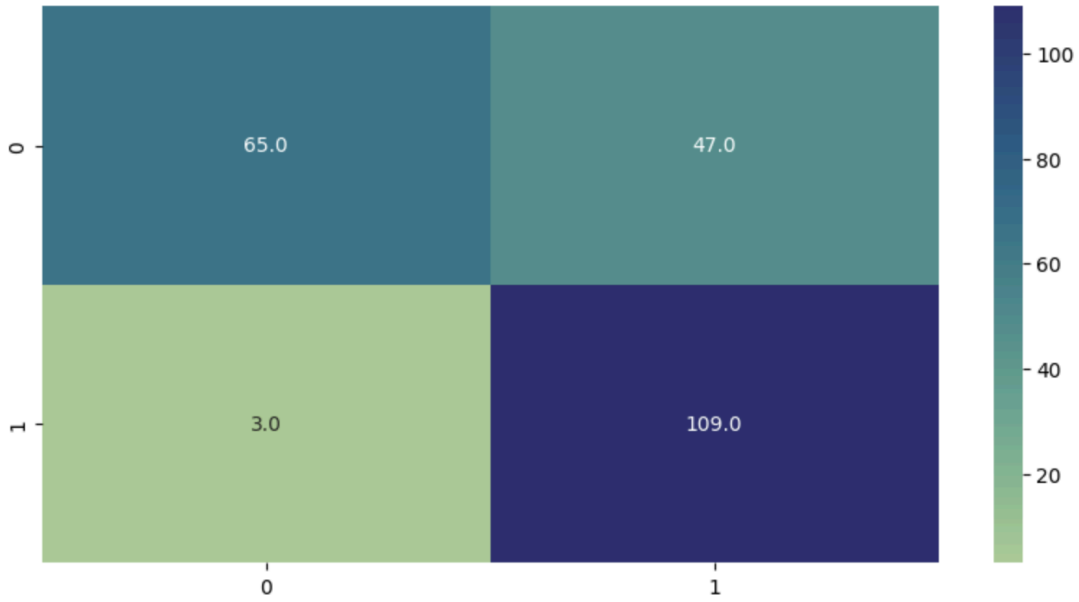
Model descriptions:

1. *VGG16:* This is a popular Convolution Neural Network (CNN) architecture that has been widely used for image classification tasks. Developed by Visual Geometry Group (VGG) at the University of Oxford, VGG16 is known for its simplicity and depth, consisting of 16 layers with learnable weights.
2. *ResNet-50:* This is a widely used Convolutional Neural Network (CNN) architecture, is a part of the Residual Networks family, which introduced the concept of residual learning. ResNet-50 is specifically designed to address the vanishing gradient problem that hampers deep neural networks, allowing them to train effectively even with very deep architectures.
3. *Efficient NetV2:* This is a state-of-the-art Convolutional Neural Network (CNN) architecture designed for efficient scaling and improved performance in image classification tasks. Developed by Google Research, Efficient NetV2 incorporates advancements in both the architecture design and training techniques, making it a powerful tool for the task like smoker detection. • Efficient NetV2 uses compound scaling to balance network depth, width and resolution. This approach ensures that the model scales efficiently without disproportionately increasing computational costs.
4. *Vision Transformer (ViT):* Vision Transformer (ViT) is a novel deep learning architecture that leverages the principles of transformers, which have revolutionized natural language processing, for image classification tasks. Unlike traditional Convolutional Neural Networks (CNNs), ViTs treat images as sequences of patches and apply transformer mechanisms to capture spatial dependencies and global context more effectively.

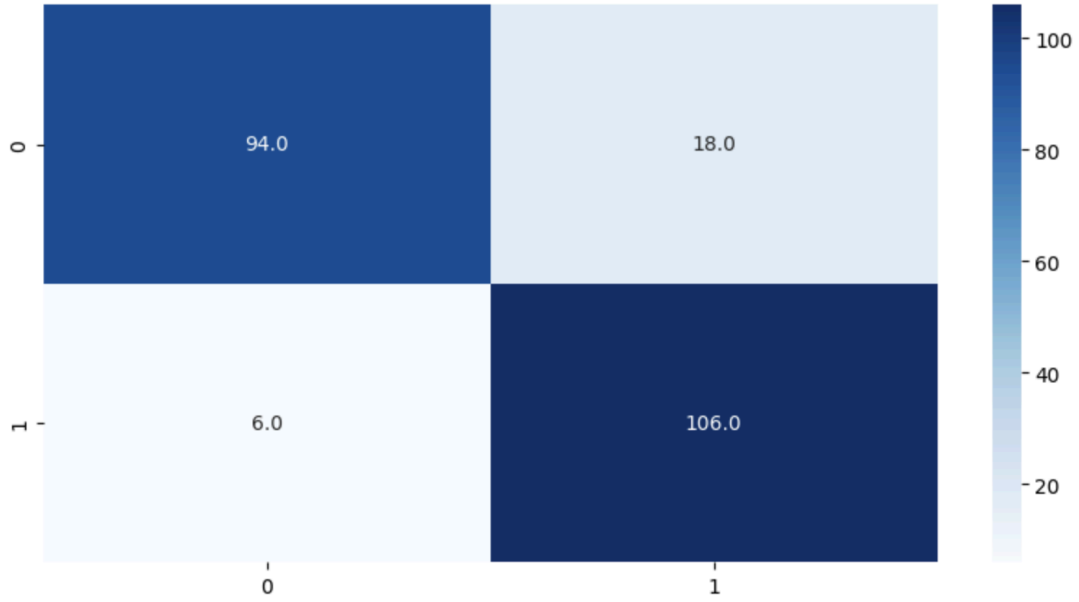
Comparative Analysis & Key Findings:

The architecture explored include VGG16, ResNet-50, Efficient NetV2, and Vision Transformer (ViT). We provide a comprehensive analysis of the performance metrics, compare the results, and discuss the implications of these findings in the context of smoker detection.

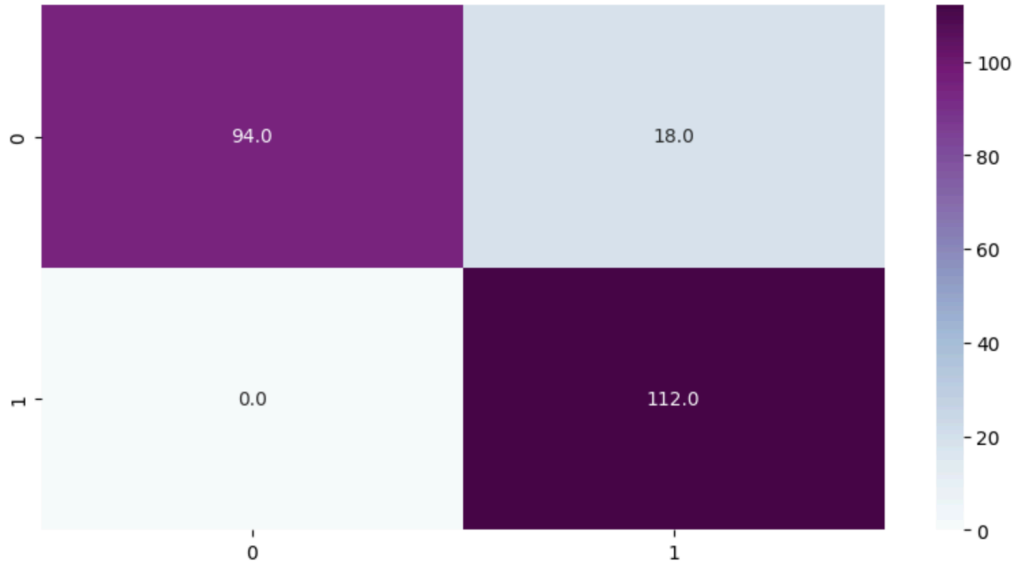
1. *Confusion matrix of VGG 16:*



2. *Confusion matrix of Resnet 50:*

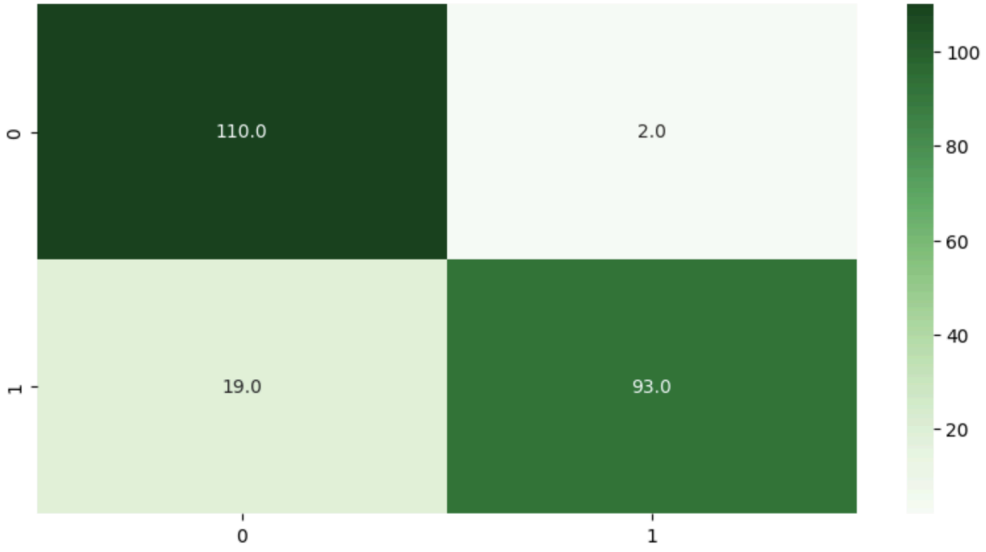


3. *Confusion matrix of Efficient Net V2:*



EfficientNetV2 outperformed ResNet-50 in all metrics, particularly in terms of recall and F1-score, suggesting it has a better balance between precision and recall. The confusion matrix demonstrates fewer misclassifications compared to ResNet-50, and confirms its superior performance.

4. *Confusion matrix of Vision Transformer (ViT):*



ViT performed slightly better than ResNet-50 but did not surpass EfficientNetV2. The confusion matrix indicates that ViT has a high true negative rate, though slightly more false negative compared to EfficientNetV2 and that's why the recall score of the Vision Transformer model is quite low.

Comparative Analysis Table:

Metric	VGG16	ResNet-50	Efficient NetV2	Vision Transformer
Accuracy	79.8%	89.29%	93.96%	91.62%
Precision	70.7%	85.48%	88.15%	97.89%
Recall	97.3%	94.64%	100%	84.04%
F1-Score	81.4%	89.83%	94.56%	90.86%
AUC	78.7%	89.29%	93.96%	91.62%

Table 1: Performance Comparison of different CNN architectures

The above result indicates that Efficient NetV2 is the most effective architecture for the task of smoker detection. Its superior performance can be attributed to its advanced design, which balanced model depth, width, and resolution more effectively than ResNet-50 and ViT.

Key Findings:

1) Efficient NetV2 achieved the highest accuracy, indicating its robustness in classifying both smokers and non-smokers accurately.
2) The balance between precision and recall in Efficient NetV2 suggests it minimizes both false positives and false negatives more effectively.
3) A higher F1-score in Efficient NetV2 confirms its overall better performance in handling classification tasks.
4) The Vision Transformer showed promising results, demonstrating the potential of transformer based models in image classification tasks traditionally dominated by CNN. However, it slightly lagged behind Efficient NetV2 in most metrics.

In summary, this project has demonstrated the effectiveness of advanced CNN architectures for smoker detection, with Efficient NetV2 standing out as the most capable model. By addressing the identified areas for further research and development, the field can continue to evolve, leading to more accurate, efficient, and ethical applications of deep learning in public health and beyond.