REGULAR ARTICLE

Huli Ka, Balbon! Automated Alarm for Smoking in Non-designated Areas Using Deep Learning

Paula Joy Martinez*†, Paulo Mario Medina†, Jeonne Joseph Ramoso† and Jeremiah Dominic Soliman†

*Correspondence:
pmartinez.msds2024@aim.edu
Aboitiz School of Innovation, Technology, and Entrepreneurship, Asian Institute of Management, Paseo de Roxas, 1229 Makati City, Philippines
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Introduction: This study presents an interpretable pre-trained model which aims to classify smoking from non-smoking images, using Xception and Grad-CAM. Alongside this, this study also aims to show that directly training the densely connected classifier on an unfrozen Xception may lead to good results.

Related Works: While previous works have discussed mostly classification techniques and the performance of classifiers on the smoking detection use case, this study aims to present a new workflow towards transfer learning and to integrate the interpretability of deep learning models applied in the given use case, using Grad-CAM.

Data and Methods: Using a Kaggle dataset on smoking and non-smoking, various models such as simple CNN, VGG-16, and Xception were evaluated using validation metrics such as accuracy, precision, and recall. Given that Xception model which is directly unfrozen performed the best, different thresholds were evaluated using test metrics. Lastly, Grad-CAM was used to identify areas within an image which served as the predictive cues for the model.

Results and Discussion: Using Xception which is directly unfrozen and with an ER Index of 0.07, the authors arrived at with a test accuracy of 96%, test precision of 95%, and test recall of 97%. Moreover, using Grad-CAM, it was found that the model focuses on the face, hands, and cylindrical objects; this also poses a limitation wherein the model may perceive any smoke-like hand gesture and stick-like object as indicative of smoking.

Conclusions and Recommendations: Directly training a densely connected classifier on the unfrozen convolutional layers of a pre-trained model was found to be feasible, especially for this use case. Moreover, in order to account for other factors such as weather and distance from the camera, it is recommended that future researchers should make use of CCTV datasets for similar smoking detection use cases.

Keywords: smoking; deep learning; pretrained models; Xception; classifier

Highlights

- Directly training a densely connected classifier to an unfrozen pre-trained model may prove feasible in acquiring optimal model results.
- Visual cues such as stick-like objects, hands, faces, and even backgrounds may account for the model's predictions and may also serve as a limitation.
- Different thresholds may be used depending on the use case where in this case, ER Index can be used to maximize recall while the baseline threshold may be used to maximize precision.

Introduction

Secondhand smoke exposure is a major public health issue affecting people of all ages. Alarmingly, around 13% of adults are exposed in workplaces, 9% in restaurants, and 12% in public transit (Campaign for Tobacco Free Kids, 2023). Even worse, over 40% of children aged 13-15 are exposed in public enclosed spaces, and nearly 30% at home (Campaign for Tobacco Free Kids, 2023). This widespread exposure puts millions at risk of serious diseases like lung cancer, heart problems, and respiratory illnesses (Arcury, et.al., 2020), and the high associated financial toll is staggering—smoking-related diseases cost the Philippines an estimated 44.6 billion pesos in 2016 alone (Global Action to End Smoking, 2024).

Clearly, effectively enforcing smoking bans in public areas is critically needed to protect people's health and reduce healthcare costs. However, current enforcement methods relying heavily on human monitoring are resource-intensive and impractical for large-scale implementation. A better approach utilizing technology-driven solutions combined with robust policies and public awareness efforts is urgently required to create smoke-free environments and safeguard public health, especially for vulnerable groups like children and adolescents.

Related Works

The topic of cigarette smoking detection has seen a recent evolution, with different approaches being employed for detection. Kavitha et al. (2017) utilized feature detection techniques like Binary Robust Invariant Scalable Keypoints (BRISK) to detect smoke in an image, which was then combined with a Haar classifier and template matching. This allowed for the detection of cigarettes within the image as well as the recording of the faces of individuals captured with detected cigarettes. Although there have been early implementations of computer vision techniques for this use case, the literature also presents sensor-based approaches. One such approach, proposed by Thakur, Poddar, and Roy (2022), involves recording and analyzing wrist movement patterns and frequencies associated with the act of smoking. Nonetheless, computer vision approaches have witnessed recent advancements due to the emergence of newer machine learning techniques, particularly deep learning networks like convolutional neural networks. Some notable works in this domain include Khan et al.'s (2022) study employing a pre-trained Inception-ResNet-V2 model for classifying smoking and non-smoking images, Aditya et al.'s (2023) work utilizing YOLOv5 for detecting cigarette smokers within images, and Zhang et al.'s (2018) research on SmokingNet, which leveraged a GoogleNet model for classifying smoking and non-smoking pictures.

While there have been recent works employing pre-trained models for this use case, the novel insights provided by this paper would be twofold. Firstly, it demonstrates a different deep learning approach by directly training a densely connected classifier with the unfrozen topmost layers of a pre-trained model, rather than training the classifier on both frozen and unfrozen components of the pre-trained model. Secondly, whereas most papers focus primarily on the classifier's performance for the given problem, this paper not only discusses the performance of the proposed model but also provides new insights into how the model classifies based on the activation areas in an image, as revealed by Gradient-weighted Class Activation Mapping (Grad-CAM).

Data and Methods

Data Source

The dataset, sourced from Kaggle (Kapadnis, 2023), consists of 1,120 candid images that is evenly split, with 50% the images labeled as "smoking" and the remaining 50% labeled as "not smoking". The 716 training images, 180 validation images, and 224 test images capture diverse scenarios. This includes people holding cigarettes (with or without visible smoke) for the "smoking" class, and various angles and zoom levels of people's faces and bodies for the "not smoking" class. All available images were utilized to provide comprehensive training data and allow for an accurate classification by the deep learning model.

Given the dataset, the Proportion Chance Criterion (PCC) is as follows, where n is the index of the class:

$$Proportion\ Chance\ Criterion = \sum_{n=0}^{Classes} \frac{m_n}{\sum_{n=0}^{Classes} m_n} \tag{1}$$

Given the equal split between the classes in the training set, the PCC can be calculated as 50%. This would then be multiplied to 1.25 to get the baseline accuracy at 62.5%.

Sample Images

Figure 1: **Smoking Images from the Training Set**
The model was trained on "smoking" images that are close-up shots. Thus, the model's performance is best for close-up shots of smokers.

Figure 1 shows training images that are labeled as "smoking". which are mostly close-up shots. It can be observed that most of the images are close-up shots and include the presence of a cigarette stick or its proximity to the mouth area. As will be discussed and shown later through the Grad-CAM heatmaps, these characteristics are the defining factors for the model's predictions.

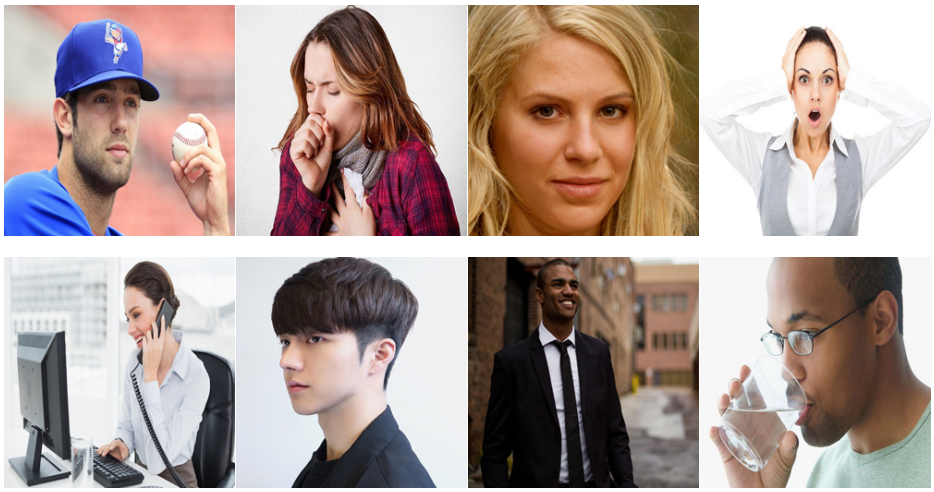


Figure 2: **Non-smoking Images from the Training Set**
The "non-smoking" images primarily consist of faces that are either very close to or relatively near the camera. Additionally, gestures similar to those associated with smoking have been included in the set.

Similarly, Figure 2 primarily features "non-smoking" images that focus on the individual's facial profile, although some images also capture the body profile. The dataset also includes a variety of gestures—such as calling, coughing, holding a ball, and drinking from a glass—to prevent the model from mistakenly associating similar hand gestures or the handling of objects with smoking.

Models

Using the Champion-Challenger method to test models (Thakku, Poddar, & Roy, 2022), where the model with the best validation accuracy is selected, the authors compiled validation metrics for various models as shown in Table 1. Both recall and precision were considered to reflect the importance of both the quantity and quality of true positives given the use case. This approach is crucial for accurately detecting smokers in an area and ensuring that those identified are indeed smokers.

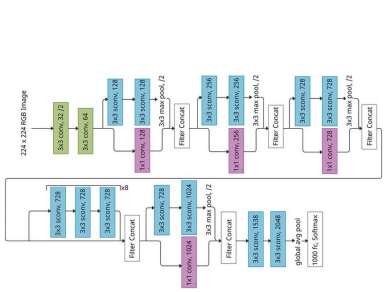
Table 1: Trained Models and their Validation Metrics

Model	Validation Accuracy	Validation Precision	Validation Recall
Raw CNN	0.72	0.69	0,78
CNN with Data Augmentation	0.52	0.57	0.18
VGG-16 (Frozen)	0.92	0.90	**0.94**
VGG-16 (Frozen then Unfrozen)	0.93	0.95	0.90
VGG-16 (Directly Unfrozen)	0.84	0.82	0.89
Xception (Frozen)	0.93	0.93	0.92
Xception (Frozen then Unfrozen)	0.92	0.92	0.91
Xception (Directly Unfrozen)	**0.96**	**0.97**	**0.94**

All the models surpassed the baseline accuracy of 62.5%. Xception (directly unfrozen) had the highest performance with 96% validation accuracy, 97% validation precision, and 94% validation recall.

The Xception model (Figure 3), as described by Chollet (2017), features 36 convolutional layers, which can be either normal or separable. Separable convolutional layers perform both nxn and 1x1 convolutions consecutively on the inputs, instead

of the single nxn convolution found in standard layers (GeeksforGeeks, 2022). Most of these convolutional layers employ a ReLU activation function either before or after the convolution process.



(a) Xception Architecture (Srinivasan, et.al., 2021)

Model: "sequential"

Layer (type)	Output Shape	Param #
rescaling (Rescaling)	(None, 250, 250, 3)	0
xception (Functional)	(None, 8, 8, 2048)	20861480
flatten (Flatten)	(None, 131072)	0
dropout (Dropout)	(None, 131072)	0
dense (Dense)	(None, 64)	8388672
dense_1 (Dense)	(None, 32)	2080
dense_2 (Dense)	(None, 1)	33

Total params: 29252265 (111.59 MB)
Trainable params: 13139585 (50.12 MB)
Non-trainable params: 16112680 (61.46 MB)

(b) Whole Model Architecture

Figure 3: **Xception and Whole Model Architectures**
The Xception architecture was added with a custom-made densely connected classifier in order to account for the use case.

Additionally, the Xception architecture consists of 14 modules, most of which incorporate linear residual connections (Chollet, 2017). These linear residual connections are convolutional layers utilizing 1x1 filters with a stride of 2 and are added to the outputs of each module (Wong, 2021). The architecture begins with 32 filters in the first module and increases in a non-discernible pattern, reaching up to 2048 filters by the final stages.

Instead of flattening the output, the model uses GlobalAveragePooling (Chollet, 2017). However, in this case, the output after the pooling process was flattened by the authors. Additionally, to adapt the model for the specific task of classifying images as smoking or not smoking, a densely connected classifier was integrated at the top of the network. This classifier includes a dropout layer with a rate of 0.5 to prevent overfitting. The densely connected segment consists of three layers: the first layer features 64 nodes with a ReLU activation function, the second layer has 32 nodes also with a ReLU activation, and the final layer contains a single node with a sigmoid activation function, which is designed to act as the binary classifier.

Lastly, the topmost layers of the model, specifically those from the 14th separable convolutional layer upwards, are unfrozen and, together with the densely connected classifier, are trained to converge immediately. This approach deviates from Chollet's (2023) recommended method, which suggests initially training the densely connected classifier with the convolutional layers completely frozen, followed by fine-tuning with the unfrozen convolutional layers. Despite this departure from Chollet's guidelines, the model demonstrated superior performance compared to the typical transfer learning fine-tuning workflow.

Thresholds for the model were also established using the Youden's Index, ER Index, and the F-1 Score (Unal, 2017). The Youden's Index is defined as the point where both the number of true positives and true negatives are maximized. This index is quantified by the following equation, where c represents the threshold variable and J denotes the Youden's index:

$$Youden's\ Index = \max_{c}(Se(c) + Sp(c) - 1) \tag{2}$$

On the other hand, the ER Index is the point where the false positives and false negatives are minimized, which nearly approximates the point nearest to the (0, 1) of the ROC curve (Unal, 2017). This is best represented by the equation:

$$ER\ Index = \min_{c}(\sqrt{((1 - Se(c))^2 + ((1 - Sp(c)))^2}) \tag{3}$$

Lastly, the F-1 score, which measures the balance between precision and recall, is used to identify the threshold where the best quantity and quality of positive classifications are achieved. The F-1 score is calculated using the following equation:

$$F - 1\ Score\ Index = \max_{c}(\frac{2 * Precision(c) * Recall(c)}{Precision(c) + Recall(c)}) \tag{4}$$

For each of these thresholds, test metrics such as accuracy, precision, and recall were calculated. Subsequently, including the base threshold of 0.5, these test metrics were compared. The threshold that maximized recall—identifying the highest number of potential positives or smokers in an image—was selected for use.

Gradient-weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping or Grad-CAM was used to find the activation areas or areas within the image which were pivotal or important to the model in its classification of images as smoker or non-smoker. According to Selvaraju, et. al. (2019), this is made possible by using the gradients of a given predicted class flowing into the final convolutional layer in order to create a heat map which shows the regions within the image important in the classification task. In the next section, these heat maps would be presented and analyzed, and the implications of the insights from the heat maps would also be discussed.

Results and Discussion

Model Predictions

Table 2: Thresholds and False Positive Rates (FPR) per Threshold

Threshold	Threshold Value	FPR
Baseline	0.5	NA
Youden's Index	1.0	0.02
ER Index	0.07	0.02
F-1 Score Index	1.0	0.02

Although there are similarities within their false positive rates, ER Index differs by a greater amount from the other indices.

As seen in Table 2, both the Youden's Index and F-1 Score Index share the same threshold, while the ER Index differs by a significant amount from the other indices. In order to fully evaluate which threshold was best for the model, both test metrics and confusion matrices were used to select the best threshold for the model.

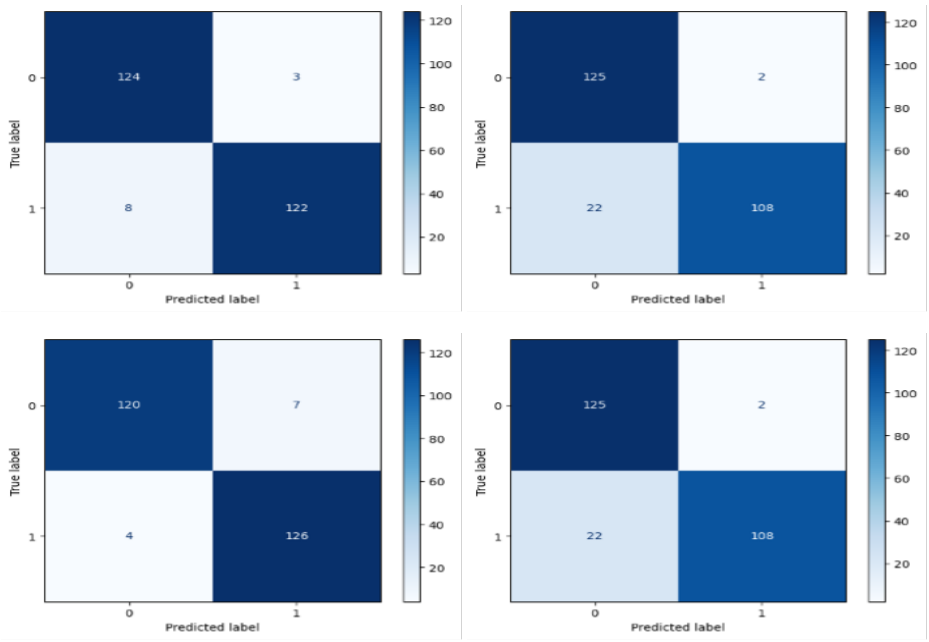


Figure 4: **Confusion Matrices per Threshold**
(Upper left: Baseline; Upper right: Youden's Index; Lower left: ER Index; Lower right: F-1 Score Index)

As seen in Figure 4, given that the false negatives or the box on the lower left of each confusion matrix is the primary number that is being minimized across models, it may be hinted that ER Index should be the threshold used for the model.

Table 3: Test Metrics per Threshold

Threshold	Test Accuracy	Test Precision	Test Recall
Baseline (0.5)	**0.96**	**0.98**	0.94
Youden's Index (1.0)	0.91	**0.98**	0.83
ER Index (0.07)	**0.96**	0.95	**0.97**
F-1 Score Index (1.0)	0.91	**0.98**	0.83

While both Youden's Index and F-1 Score Index performs worse as compared to the other two thresholds, it can be said that the other two thresholds, the baseline threshold and ER Index, have differing purposes on how they could be used.

Table 4: Final Test Metrics

Test Accuracy	Test Precision	Test Recall
0.96	0.95	0.97

The final test metrics for the final model would be test accuracy of 0.96, test precision of 0.95, and test recall of 0.97.

Gradient-weighted Class Activation Mapping

As illustrated in Figure 5, the heat map generated by Grad-CAM shows that the red areas, which indicate regions of high activation, are predominantly highlighted on the face and the object held by the person. Notably, in the third image, the model focused not on the face but on the presence of a stick-like object in the person's hand, leading it to predict that the image contained a smoker. This suggests that the model associates certain object shapes with smoking behavior.

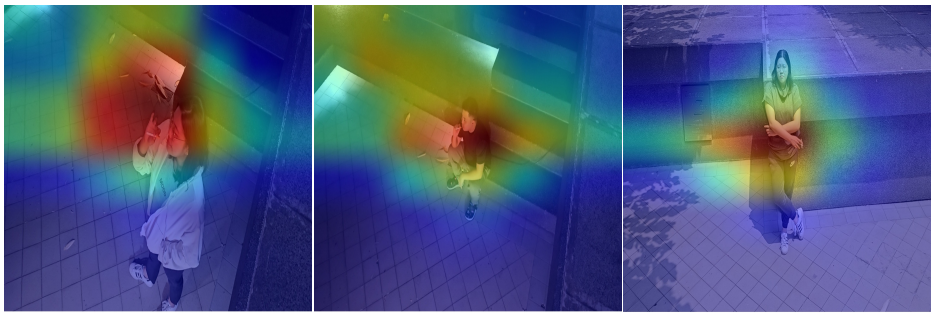


Figure 5: **GradCam Heatmaps on True Smoking Images**
It can be said that in predicting an image to containing a smoker, the model gives more attention to the body, face, and object that the person is holding.

Figure 6: **GradCam Heatmaps on True Non-smoking Images**
It can be said that in predicting an image to not containing a smoker, the model could either give more attention to the face or to the background.

For non-smoking images, as shown in Figure 6, the model's attention—indicated by the heat map—may focus on the hand, face, or background. It can be inferred that the heat observed in the background of the second image is likely due to the presence of a stick-like design. This suggests that the model may sometimes misinterpret background elements with shapes similar to cigarettes as relevant cues. This indicates potential areas for further refinement in distinguishing relevant from irrelevant features in the classification process.

As shown in Figure 7, images featuring gestures and objects that resemble smoking but do not actually depict smoking were misclassified as smoking. This misclassification may arise from the model's tendency to generalize all cylindrical objects as cigarette sticks and to interpret hand gestures typically associated with smoking as indicative of smoking behavior, even in the absence of a cigarette. This highlights a limitation of the model: its overgeneralization of cylindrical objects as cigarette sticks and smoke-like gestures as smoking, which can lead to inaccuracies in classification.

Included along with the article is a smoker-detection application that utilizes the model and the ER Index threshold. This application is with and without the Grad-CAM filter to provide users with the option to observe how the model's attention affects its predictions.

Figure 7: **GradCam Heatmaps on "Smoking" Images**
Objects such as pens and handcreams which were placed near the mouth or near the body caused the model to misclassify these images as smoking.

Conclusion and Recommendations

Conclusion

With a test accuracy of 96%, test precision of 95%, and test recall of 97%, the best model is Xception with an ER Index threshold of 0.07. The methodology employed in developing this model deviates from the standard approach to fine-tuning pre-trained models. Typically, the densely connected classifier is first trained with the convolutional layers frozen to stabilize the learned features before fine-tuning. However, in this case, the classifier was directly trained with the unfrozen convolutional layers, which allowed for simultaneous updates across all layers.

Additionally, insights gained from using Grad-CAM revealed specific focus areas of the model in classifying images. When identifying smoking-related images, the model primarily concentrates on the face and any cylindrical or stick-like objects held by individuals. For non-smoking images, the model extends its focus to include not just these elements but also the background.

A notable limitation of the model, however, is its tendency to misclassify images where non-smoking subjects exhibit gestures or hold objects that mimic smoking behavior. This overgeneralization, where any cylindrical object or smoke-like gesture is deemed indicative of smoking, underscores a critical area for improvement in ensuring more accurate and context-aware classifications.

Recommendations

The current model, trained solely on candid images of individuals either smoking or not smoking, would benefit greatly from incorporating real CCTV footage into the dataset. Since CCTVs operate around the clock, the footage would capture diverse lighting conditions, which is a crucial factor for enhancing the model's robustness and ability to handle real-world scenarios accurately. Additionally, data augmentation techniques could be employed to artificially introduce variations in lighting, angles, and other factors that could further improve the model's generalization capabilities. Moreover, the model should be trained to handle multiple people in a single frame, as CCTV cameras are typically placed in crowded environments. By implementing these recommendations, the model's accuracy and reliability in detecting smoking behavior from CCTV footage could be significantly improved and better aligned with the desired operational requirements for deployment.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
PJM was crucial in shaping the overall narrative of the study and in the continuous enhancement of the model's architecture. PMM and JDS were extensively involved in the practical implementation and operationalization of both the model and the *Smokeception Cam* application. JJR primarily focused on analyzing and cross-validating the results. All authors contributed significantly to the conceptualization, execution, analysis, and the collaborative writing of this paper.

Acknowledgements
The authors would like to express their sincere gratitude to their mentors, Dr. Christopher Monterola, Prof. Kristine Ann Carandang, and Prof. Leodegario Lorenzo. Their guidance was invaluable in refining the narrative of the study and continuously improving of the model.

References
1. Arcury, T., et al. (2020, November). "It's worse to breathe it than to smoke it": Secondhand smoke beliefs in a group of Mexican and Central American immigrants in the United States. *International Journal of Environmental Research and Public Health, 17*(22), 8630. Retrieved from: https://doi.org/10.3390/ijerph17228630
2. Aditya, M., et al. (2023, February 11). Smoking detection using deep learning. *International Journal of Computer Trends and Technology, 17*(2), 8-14. Retrieved from https://ijcttjournal.org/2023/Volume-71
3. Avram, R., Tison, G. H., Aschbacher, K., Kuhar, P., Vittinghoff, E., Butzner, M., Runge, R., Wu, N., Pletcher, M. J., Marcus, G. M., & Olgin, J. (2019). Real-world heart rate norms in the Health eHeart study. *NPJ digital medicine, 2,* 58. https://doi.org/10.1038/s41746-019-0134-9
4. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251-1258, https://arxiv.org/pdf/1610.02357.pdf
5. Chollet, F. (25, June 2023). *Transfer learning & fine-tuning*. Keras. Retrieved from: https://keras.io/guides/transfer_learning/?fbclid=IwAR0jdMWxBu2BD6ysPC_A0ZQooKyn67V0d4b1-MxTUVYP3zUCBAd1kR5IJdA
6. GeeksforGeeks. (29, September 2022). *Depth wise Separable Convolutional Neural Networks*. https://www.geeksforgeeks.org/depth-wise-separable-convolutional-neural-networks/
7. Global Action to End Smoking. (n.d.). *State of Smoking and Health in the Philippines.* Retrieved from: https://globalactiontoendsmoking.org/research/tobacco-around-the-world/philippines/?fbclid=IwAR1SWncmw-gcOiqJ3MCZ9ZQeNLzFntOlnoYTIhG8DYKsHSZJ4decSqhbiDk
8. Kapadnis, S. (n.d.). *Smoker Detection [Image] classification Dataset*. Kaggle. Retrieved from: https://www.kaggle.com/datasets/sujaykapadnis/smoking?select=Training&fbclid=IwAR33eRiInvXXjicNmyM_MwFs12bnOozWhJcZYbjmnwW6ajH6L3fGQETPfw
9. Kavitha, H., et al. (2017). Automated smoking for smoking detection. *International Journal of Current Engineering and Scientific Research (IJCESR), 4*(4), pp. 64-67. Retrieved from: https://troindia.in/journal/ijcesr/vol4iss4/64-67.pdf
10. Khan, A., et al. (2022, January 24). CNN-based smoker classification and detection in smart city application. *Sensors, 22*(3), 892. https://doi.org/10.3390/s22030892
11. Srinivasan, K. et al. (May 2021). Performance Comparison of Deep CNN Models for Detecting Driver's Distraction. *Computers, Materials & Continua, 68*(3), pp. 4109-4124. Retrieved from: https://doi.org/10.32604/cmc.2021.016736
12. Tobacco Free Kids. (16, October 2023). *The Toll of Tobacco in the Philippines*. Retrieved from: https://www.tobaccofreekids.org/problem/toll-global/asia/philippines?fbclid=IwAR2LE7hFhhKGDHw4wZyMNkh_W-LoLiRgiNP7by6EkYdjldWfD2M7tQ2F4L4
13. Unal I. (2017). Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. *Computational and mathematical methods in medicine,* 2017, 3762651. https://doi.org/10.1155/2017/3762651
14. Wong, W. (19, December 2021). *What is Residual Connection? A technique for training very deep neural networks*. Towards Data Science. Retrieved from: https://towardsdatascience.com/what-is-residual-connection-efb07cab0d55
15. Selvaraju, R., et al. (2019, December 3). *Grad-CAM: Visual explanations from deep networks via gradient-based localization*. arXiv. https://arxiv.org/abs/1610.02391v4
16. Thakur, S., Poddar, P., & Roy, R. (2022, February 25). Real-time prediction of smoking activity using machine learning-based multi-class classification model. *Multimedia Tools and Applications, 81,* pp. 14529–14551. https://doi.org/10.1007/s11042-022-12349-6
17. Wang, Z., et al. (2023, August 24). Smoking behavior detection algorithm based on YOLOv8-MNC. *Frontiers in Computational Neuroscience, 17,* https://doi.org/10.3389/fncom.2023.1243779

Additional Files
Additional file 1 — ML3_LT1_Appendix.ipynb
This notebook contains the overall pipeline and appendix of the study (with a widget to see Grad-CAM in action with the images).

Additional file 2 — /Application

This folder contains the smoker detection application, with both the Grad-CAM filter and without the filter.