Leveraging Lightweight Design and Attention for Lung Disease Predictions from CXR Images

Andrea Marinelli†

***Abstract*—The COVID-19 pandemic has accentuated the need for rapid and accurate diagnostic methodologies, challenging the capacities of traditional medical imaging techniques. This study explores the application of lightweight Convolutional Neural Networks (CNNs) enhanced with Squeeze-and-Excitation (SE) blocks for an efficient and accurate diagnosis of lung disease from chest X-ray (CXR) images. We propose novel architectures integrating these elements, offering a dual advantage of high diagnostic performance and reduced computational cost. Extensive evaluations on a diverse dataset demonstrate that lightweight models achieve competitive accuracy in this domain, particularly when enhanced with channel attention, offering a viable solution for resource-constrained environments. The integration Class Activation Mapping (CAM) further allows to visualize the relevant image regions that the models focus on, adding a layer of interpretability to the automated decision-making process. These techniques could potentially streamline the diagnostic workflow in clinical settings, making rapid, accurate screening accessible in resource-limited situations.**

***Index Terms*—Deep Learning, Medical Imaging, Convolutional Neural Networks, Chest X-rays, Lightweight Models, Attention Mechanisms.**

I. Introduction

The COVID-19 pandemic underscored the critical need for rapid and accurate screening methods to detect infected patients, highlighting the limitations of traditional approaches. The primary method used for detecting COVID-19 cases is reverse transcriptase-polymerase chain reaction (RT-PCR) testing, which is highly specific in identifying SARS-CoV-2 RNA from respiratory samples [1]. However, RT-PCR is time-consuming, resource-intensive, and requires expert interpretation, prompting the exploration of alternative, more accessible methods.

One such alternative is chest X-ray (CXR) imaging, which has become a vital tool due to its widespread availability, rapid acquisition, and lower cost compared to other diagnostic techniques. CXR images are commonly used in healthcare settings to identify lung abnormalities. However, interpreting these images can be challenging because the visual signs of diseases like COVID-19 or pneumonia can be subtle, often necessitating the expertise of experienced radiologists. This creates a bottleneck in critical situations where swift diagnosis is essential.

†Department of Mathematics "Tullio Levi-Civita", University of Padova, email: {andrea.marinelli}@studenti.unipd.it

To overcome these challenges, recent advances in deep learning have shown promise in automating the analysis of CXR images. Convolutional neural networks (CNNs) can process large datasets, learning to recognize patterns that may not be easily detectable by human observers. These models can assist radiologists by highlighting areas of concern or even enabling fully automated diagnoses, thus improving both the speed and accuracy of lung disease detection.

Despite significant progress, much of the existing research has focused on improving model accuracy, often overlooking the importance of computational efficiency. In resource-constrained environments, it is crucial to develop lightweight models that maintain high accuracy while being computationally affordable. Therefore, with this study, we aim to contribute to addressing this gap by seeking to answer the following research questions:
Q1: Can a lightweight model with a low number of parameters and low computational cost still perform well in this domain?
Q2: Can incorporating soft attention mechanisms, such as SE blocks, help these models achieve better performance?
To this end, we have explored lightweight design principles, including depthwise and pointwise convolutions, inverted residual blocks, and soft attention mechanisms like squeeze-and-excitation (SE), which have shown promise in improving model performance across various tasks. Overall, the contributions of this study can be summarized as follows:

- We propose lightweight CNN architectures that maintain competitive performance with a low computational cost.
- We explore the effectiveness of SE blocks in enhancing model performance for medical imaging applications.
- We provide insights into the trade-offs between accuracy and efficiency in deep learning models for CXR analysis.

This paper is structured as follows: Section II reviews the related work, Section III describes the system and data models, Section IV details the proposed methodology, Section V presents the experimental results, and Section VI concludes with final remarks.

II. Related Work

This section reviews key developments in deep learning for chest X-ray analysis, lightweight neural network designs, and the use of attention mechanisms to enhance interpretability and performance in medical imaging.

a) Deep Learning for Chest X-ray Images: The problem of chest X-ray image classification has been extensively explored in the field of medical image analysis. In 2017, Rajpurkar et al. [2] introduced an algorithm based on DenseNet that outperformed practicing radiologists in detecting pneumonia from chest X-rays, demonstrating the potential of deep learning for lung disease detection. Following the release of open-source datasets such as the COVID-19 Image Data Collection by Cohen et al. [3], numerous studies have been conducted on COVID-19 detection using CXR images. Among these, COVID-Net by Wang et al. [1] is particularly noteworthy, as it introduced a tailored architecture for the detection of COVID-19 using generative synthesis, driven by the need for faster interpretation of CXR images.

b) Lightweight Design: The development of lightweight neural network architectures has been crucial for deploying deep learning models in resource-constrained environments. MobileNetV1, introduced by Howard et al. in 2017 [4], employed depthwise separable convolutions to drastically reduce computational cost and model size. This line of work continued with MobileNetV2 and MobileNetV3, which introduced further optimizations such as inverted residuals and squeeze-and-excitation (SE) blocks, achieving an improved balance between efficiency and accuracy [5], [6]. Other notable lightweight models include ShuffleNet [7] and EfficientNet [8], both of which introduced innovative strategies for optimizing the trade-off between model performance and computational demands.

c) Visual Attention for Chest X-ray Images: Visual attention mechanisms have emerged as powerful tools in deep learning, allowing models to focus on specific regions of an image, thereby enhancing performance and providing interpretability in visual data analysis. This latter capability is particularly important in medical imaging, where understanding whether a model is focusing on clinically relevant areas is crucial for trust in automated diagnosis. One widely used technique in this domain is Class Activation Mapping (CAM), introduced by Zhou et al., which allows for the visualization of the regions within an image that contribute most to the model's prediction [9]. CAM has been instrumental in verifying that deep learning models for chest X-rays attend to relevant anatomical structures when diagnosing diseases. For diseases that present in localized areas, such as certain lung nodules, hard attention mechanisms can be beneficial by allowing the model to selectively focus on the most clinically significant regions. However, diseases like pneumonia and COVID-19 often manifest with more diffuse patterns, including ground-glass opacities, consolidations, and pleural effusions. These abnormalities, characterized by varying levels of brightness and opacity, are well-suited to detection through channel attention mechanisms, which fall under the category of soft attention. These mechanisms involve assigning adaptive weights to different channels of the feature maps, thereby emphasizing the most relevant features across the image. An example of channel attention is the aforementioned Squeeze-and-Excitation Blocks [10]. This enhances the model's ability to distinguish between subtle differences in chest X-ray images, improving diagnostic accuracy.

The work presented in this paper builds upon these foundations by proposing models that integrate lightweight architectures with attention mechanisms, aiming to optimize both computational efficiency and diagnostic accuracy in chest X-ray analysis.

III. Processing Pipeline

In this section we delve more into our work, following the pipeline used for the experiments. We aim to introduce, at high level, the methods and approach employed in the study. Our focus was on exploring efficient CNN architectures, specifically focusing on models from the MobileNet family. We explored the possibility to enhance these baseline models by incorporating channel attention mechanisms, to boost the performance. In the following subsections, we describe the baseline models, explain how they were enhanced with attention mechanisms, and how they were evaluated and compared. Finally, we highlight how attention mechanisms were used to visualize the areas to attend when classifying the disease.

a) Baseline: We chose MobileNetV1 and MobileNetV2 as our baseline models due to their innovative approaches to lightweight model design, which ensures a good balance between computational efficiency and accuracy. After carefully reviewing their key features and architectural components, we implemented these models from scratch using the Keras TensorFlow API. We varied the width multiplier α in our experiments to explore how changes in model size affect performance.

b) SE-enhanced architectures: To further enhance the architecture, we integrated Squeeze-and-Excitation (SE) blocks into both MobileNetV1 and MobileNetV2. SE blocks are a type of attention mechanism that adaptively recalibrates channel-wise feature responses, allowing the model to focus on the most relevant features within the chest X-ray images. This integration aimed to improve the models' ability to detect subtle patterns associated with diseases like COVID-19 and pneumonia, which are often challenging to identify due to their diffuse manifestations in lung tissue.

c) Evaluation and comparison: Following the implementation of the SE-enhanced architectures, we systematically evaluated the models across various configurations, again varying the width multiplier α to assess the impact on both performance and computational cost. The models were then compared in terms of performance and complexity, providing a comprehensive understanding

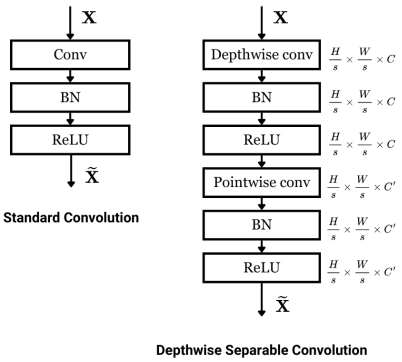

Fig. 1: Left: Standard convolutional layer with batchnorm and ReLU. Right: Depthwise Separable convolutions with Depthwise and Pointwise layers followed by batchnorm and ReLU.

of the trade-offs involved in lightweight model design for medical image analysis.

d) Class activation visualization: In addition to performance evaluation, we employed visualization techniques, specifically Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM), to generate activation maps. These maps highlight the regions of the chest X-ray images that the models focused on during the classification process, thereby enhancing the interpretability of the results and building trust in the automated diagnosis process.

In summary, our processing pipeline integrates lightweight design principles with advanced attention mechanisms, providing a robust framework for efficient and interpretable lung disease classification. The combination of MobileNet-based architectures, SE blocks, and activation map visualization forms the foundation of our approach, guiding the development of models that are both computationally efficient and diagnostically accurate.

IV. Architectures

In this section, we provide a comprehensive analysis of the architectural components and attention mechanisms employed in this work.

A. MobileNetV1

MobileNetV1 is a lightweight convolutional neural network architecture specifically designed for mobile and embedded vision applications. The key innovation in MobileNetV1 is the introduction of depthwise separable convolutions, which significantly reduce the number of parameters and computational cost compared to traditional convolutional networks, making it ideal for resource-constrained environments.

The architecture of MobileNetV1 consists of a standard convolution layer followed by multiple layers of depthwise

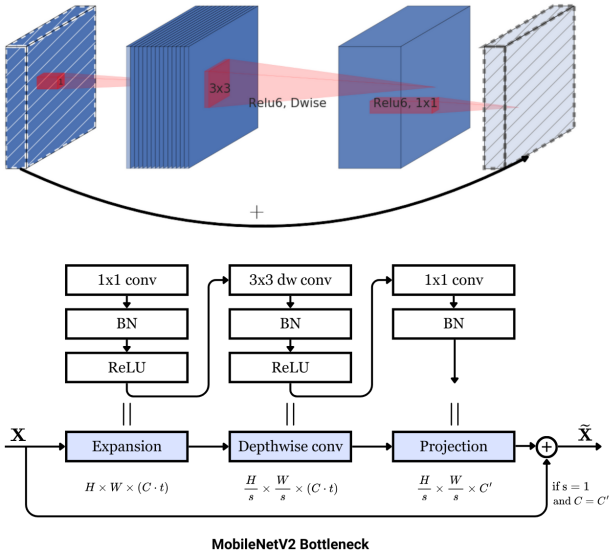

Fig. 2: Bottleneck residual block transforming from C to C' channels, with stride s, and expansion factor t. Diagonally hatched layers do not use non-linearities. The thickness of each block indicate its relative number of channels.

separable convolutions. Each depthwise separable convolution is composed of a depthwise convolution followed by a pointwise convolution (a 1x1 convolution). This design splits the convolution operation into two smaller operations, reducing computational complexity while maintaining accuracy.

The depthwise separable convolution block in MobileNetV1 is composed of two primary operations, as we can see in Fig. 1:

- **Depthwise Convolution**: This operation applies a single convolutional filter to each input channel separately, performing a spatial convolution on each channel independently. This drastically reduces the computational load compared to a standard convolution that mixes channels.
- **Pointwise Convolution**: Following the depthwise convolution, a 1x1 pointwise convolution is applied to combine the outputs of the depthwise convolution. This step is essential for mixing the information across different channels.

MobileNetV1 introduces two hyperparameters to further reduce the model size and computation:

- **Width Multiplier** (α): This parameter controls the number of channels in each layer. By scaling the number of channels by α, the computational cost and the number of parameters can be reduced proportionally.
- **Resolution Multiplier** (ρ): This parameter scales the input resolution of the image. By decreasing the input image size by ρ, the computational demand is reduced as well.

B. MobileNetV2

MobileNetV2 builds upon the principles of MobileNetV1 but introduces several key improvements. The core idea in MobileNetV2 is the introduction of the inverted residual block with linear bottlenecks, which further reduces the computational cost while preserving model accuracy.

The architecture of MobileNetV2 follows a similar structure to MobileNetV1 but replaces the depthwise separable convolution blocks with inverted residual blocks. This module is designed to efficiently capture features with fewer parameters and consists of three main components, as shown in Fig. 2:

- **Expansion Layer**: The first 1x1 convolution expands the input dimensions by a factor t, the expansion factor. This creates a higher-dimensional space where the depthwise convolution can capture more complex features.
- **Depthwise Convolution**: Similar to MobileNetV1, this operation is applied to each channel separately, maintaining the benefits of reduced computation.
- **Projection Layer**: The final 1x1 convolution projects the high-dimensional data back to a lower-dimensional space, serving as the bottleneck. If the input and output dimensions are the same and the stride is 1, a residual connection is added to facilitate gradient flow and improve learning.

The expansion factor t is a crucial parameter in MobileNetV2 that controls the dimensionality expansion in the bottleneck block. A higher t value allows the network to capture more complex features, but also increases the computational load.

C. Squeeze-and-Excitation (SE) Block

The SE block enhances the representational power of a network by explicitly modeling the interdependencies between the channels of its convolutional features. By selectively weighting the importance of different channels, SE blocks allow the network to focus more on informative features while suppressing less useful ones, improving overall model performance, particularly in challenging tasks such as medical image analysis.

The SE block operates in two stages:

- **Squeeze**: Global average pooling is applied to each channel of the convolutional feature map, producing a channel descriptor that captures the global distribution of each feature.
- **Excitation**: These descriptors are then passed through a small fully connected network (typically a bottleneck architecture with a reduction ratio) that outputs channel-wise weights. These weights are applied to the original feature map, effectively recalibrating the importance of each channel.

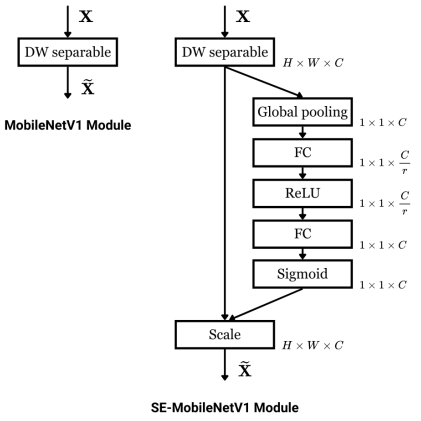

Fig. 3: The schema of the original Deptwise Separable Convolution Module in MobileNetV1 (left) and the SE-MobileNetV1 module (right).

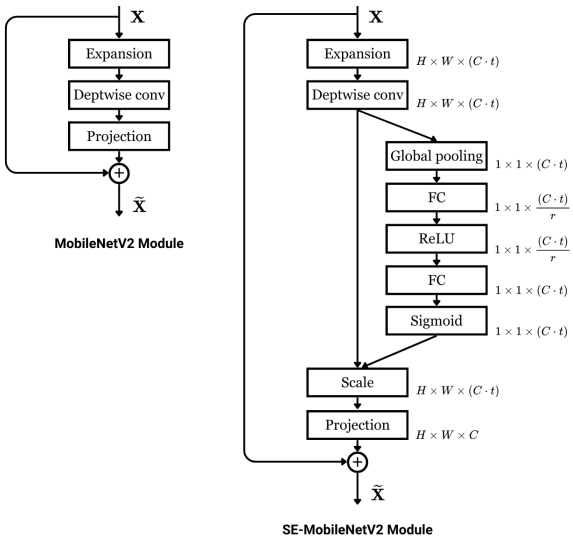

Fig. 4: The schema of the original Bottleneck Module in MobileNetV2 module (left) and the SE-MobileNetV2 module (right).

In our work, SE blocks were integrated into both MobileNetV1 and MobileNetV2 architectures to enhance their ability to focus on relevant features. The insertion of SE blocks was done selectively to maintain a balance between computational complexity and performance.

In MobileNetV1, SE blocks were added after the depthwise separable convolution, as shown in Fig. 3, following the approach presented in Hu et al. [10]. This method treats the depthwise separable block as a single unit and applies the SE modulation to the entire output of the block. Additionally, we explored another approach where the SE block was inserted after the depthwise convolution but before the pointwise convolution (the 1x1 convolution) to selectively modulate the channels before they are combined through the pointwise

convolution. However, the first approach proved more suitable and resulted in significantly better performance.

In MobileNetV2, SE blocks were added after the depthwise convolution and before the final projection layer within the bottleneck block, as shown in Fig. 4, following the approach proposed in Hu et al. [10] for residual networks.

We selectively inserted SE blocks into convolutional blocks that process features at reduced spatial scales, particularly those with a spatial resolution equal to or smaller than 14x14. This decision was inspired by the design of MobileNetV3 [6] and based on the observation that SE blocks are most effective in deeper layers, where high-level features are formed, and in blocks with a larger number of filters, where selective channel weighting can have a more significant impact.

Future experiments could explore the incremental addition of SE blocks, introducing them step by step from deeper to shallower layers while monitoring the trade-offs between accuracy and computational complexity. This approach would allow for more precise optimization of the network architecture, ensuring that SE blocks are employed where they provide the greatest benefit.

V. Data

This study used the "COVID19, Pneumonia and Normal Chest X-ray PA Dataset" released by Asraf et al. [11], which includes 4,575 posteroanterior (PA) chest X-ray images categorized into 613 COVID-19 cases, 1,525 pneumonia cases, and 1,525 normal cases. The COVID-19 images were sourced from various repositories including GitHub [3], Radiopaedia [12], TCIA [13], SIRM [14], and supplemented by 912 augmented images from Mendeley [15]. Images of pneumonia and normal cases were obtained from the Kaggle repository [16] and the NIH dataset [17]. A balanced and diverse dataset is essential for effectively training deep learning models, as it helps the model generalize better across different cases. However, the integration of data from various sources introduced challenges in processing the images.

a) Handling Diverse Image Formats: Initial exploration revealed differences in image shapes and formats, necessitating specific adjustments for managing these discrepancies. To identify the formats of the images, the `filetype` library was employed, which detected images encoded in .jpg, .png, and .bmp formats. It failed to recognize the format of 14 images, which were marked as corrupt and subsequently removed from the dataset during the loading phase as they were not usable. In the dataset creation pipeline, the `decode_image` function from `tensorflow.io` was utilized to properly handle the decoding of images in the three identified formats.

b) Preprocessing: To ensure uniform input dimensions for the model, all images were resized to 224x224 pixels. Normalization was applied by dividing pixel values by 255. Additionally, to mitigate the impact of frequently embedded textual information in chest X-ray images, the top and bottom

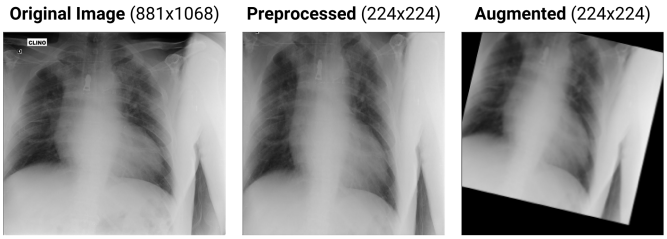

Fig. 5: From left to right: the original image, the image after preprocessing (including resizing, cropping and normalization), and the image after augmentation techniques.

8% of each image was cropped prior to training, as in Wang et al. [1].

c) Data Augmentation: To enhance the robustness of the deep neural network architectures tested, data augmentation techniques inspired by Wang et al. [1] were employed, including zoom ($\pm 15\%$), translation ($\pm 10\%$ in both x and y directions), rotation (± 10 degrees), horizontal flip, and intensity shift ($\pm 10\%$).

Moreover, the dataset creation pipeline incorporated several techniques to enhance model training efficiency and performance. A caching system was employed after preprocessing to accelerate access during training. Shuffling was utilized to randomize the order of images, thus preventing the model from learning unintended biases. Images were grouped into batches to allow for more efficient computation and gradient updates during training. Prefetching was used to prepare data batches while the model was training on the current batch, reducing idle time and improving throughput.

VI. Experiments

The experiments were conducted on Google Colab using an 8GB GPU, following the training protocol outlined by Rajpurkar et al. (2017) [2]. The network was trained end-to-end using the Adam optimizer with standard parameters: $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as recommended by Kingma Ba (2014). We used mini-batches of size 16, and fixed the maximum number of epochs to 80.

To prevent overfitting, we applied EarlyStopping, monitoring the validation loss and stopped training if no improvement was observed for 10 consecutive epochs, starting from epoch 20. This allowed the model to explore the parameter space without overfitting.

We started with an initial learning rate of 0.01. To ensure stable model convergence, we applied a learning rate reduction strategy. Specifically, when the validation loss stopped improving for 5 consecutive epochs, we decreased the learning rate by a factor of 10. This adaptive approach helped achieve more consistent progress during training.

A. Learning Rate Grid Search

To determine the optimal learning rate, we performed a grid search, training the MobileNetV1 model for 30 epochs with

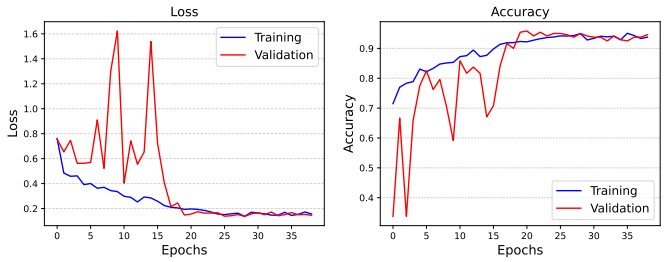

Fig. 6: Loss (left) and accuracy (right) training curves of MobileNetV1.

three different learning rates: 1×10^{-2}, 1×10^{-3}, and 1×10^{-4}. The results of this grid search indicated that a learning rate of 0.01 provided the best balance between training speed and stability.

As shown in Figure 6, this high learning rate caused the learning process to be somewhat unstable during the initial epochs, characterized by fluctuations in the validation loss. However, it facilitated rapid exploration of the loss surface, enabling the model to quickly escape from suboptimal regions and settle into a better local minimum. This behavior is advantageous in the early stages of training, where large updates can help overcome poor initializations and lead to faster convergence.

Once the validation loss began to stabilize, the learning rate reduction strategy ensured that the updates became more fine-grained, allowing the model to refine its parameters and achieve better generalization performance. This combination of an initially high learning rate followed by systematic reductions proved effective in balancing the need for rapid learning with the requirement for stable convergence.

VII. Results

In this section, we report the performance of our lightweight models on the Chest X-Ray dataset. We first evaluate the baseline models, MobileNetV1 and MobileNetV2, and then compare them with their enhanced versions incorporating channel attention mechanisms. As outlined in Section III, we explore the trade-offs between accuracy and computational cost across different model configurations.

The evaluation metrics, shown in Table 1, include Top-1 Accuracy, Precision, Recall, and F1-Score for assessing model performance, as well as the number of parameters and Mult-Adds to measure model complexity. We also present class activation maps to visualize how the models focus on different regions of the X-rays during classification.

A. Baseline Comparison: MobileNetV1 vs. MobileNetV2

Fig. 7 show the performance of the baseline MobileNetV1 and MobileNetV2 models across a spectrum of different model sizes. MobileNetV1 consistently outperformed MobileNetV2 in all configurations, which was unexpected. Although MobileNetV2 introduced innovations like inverted residuals and linear bottlenecks aimed at improve efficiency,

Network	Top-1	Prec	Rec	F1	Params	MAdds
1.0-MobileNetV1	97.4	97.4	94.8	96.1	3.2M	575.3M
0.75-MobileNetV1	96.3	95.4	93.5	91.6	1.8M	330.4M
0.5-MobileNetV1	95.7	95.3	91.6	93.0	0.8M	152.8M
0.25-MobileNetV1	96.8	96.1	94.2	93.6	0.2M	42.7M
1.0-MobileNetV2	95.0	91.8	93.5	93.5	2.3M	323.0M
0.75-MobileNetV2	94.2	93.8	88.4	88.8	1.3M	189.6M
0.5-MobileNetV2	92.9	98.4	80.0	84.9	0.6M	91.3M
0.25-MobileNetV2	91.6	93.9	80.0	82.3	0.2M	28.1M
1.0-SE-MobileNetV1	97.6	97.4	95.5	95.5	3.7M	576.5M
0.75-SE-MobileNetV1	97.8	98.0	95.5	96.9	2.1M	331.1M
0.5-SE-MobileNetV1	97.4	99.3	92.9	96.2	1.0M	153.3M
0.25-SE-MobileNetV1	96.3	93.1	96.1	92.7	0.3M	42.9M
1.0-SE-MobileNetV2	96.1	92.5	96.1	93.6	2.8M	444.5M
0.75-SE-MobileNetV2	94.6	93.9	89.7	90.3	1.6M	262.6M
0.5-SE-MobileNetV2	96.3	93.1	96.1	93.7	0.7M	127.9M
0.25-SE-MobileNetV2	91.6	83.0	94.2	86.7	0.2M	40.3M

TABLE 1: Comparison of MobileNet models with accuracy, precision, recall, f1-score, num_params, and mult-adds.

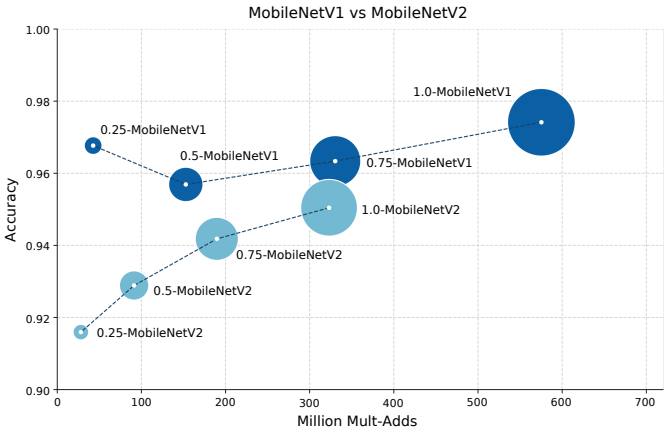

Fig. 7: MobileNetV1 vs MobileNetV2.

these changes did not translate into better performance in this domain.

For instance, at $\alpha = 1.0$, MobileNetV1 achieved 97.4% accuracy, compared to 95.0% for MobileNetV2. This performance gap remains across other values of α, with MobileNetV1 maintaining better results, even when both models have similar parameters and computational costs (MAdds). This suggests that MobileNetV2's architectural enhancements may not be as effective for lung disease classification, where MobileNetV1's simpler design with depthwise separable convolutions seems better suited.

B. SE-Enhanced Models: Performance Gains with Channel Attention

We next evaluated the effect of adding squeeze-and-excitation (SE) blocks to the architectures. As shown in Fig. 8, SE blocks led to notable performance improvements, particularly for $\alpha = 0.5$ and $\alpha = 0.75$. In these configurations, the SE-enhanced versions of both MobileNetV1 and MobileNetV2 showed significantly higher F1-scores and recall compared to the baseline models.

For example, SE-MobileNetV1 with $\alpha = 0.75$ achieved an F1-score of 96.9%, compared to 91.6% for the non-SE version, highlighting how channel attention can help the model

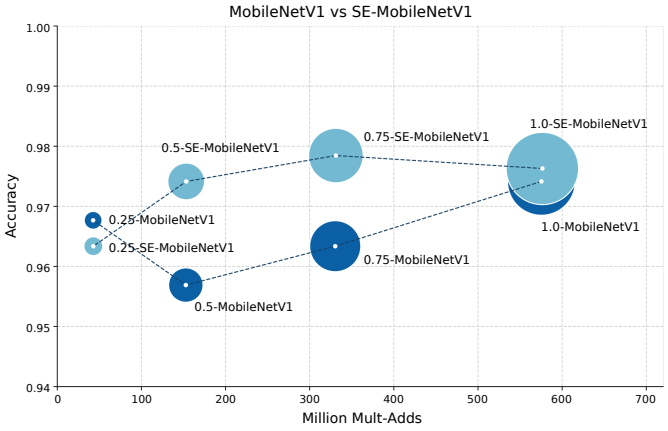

Fig. 8: MobileNetV1 vs SE-MobileNetV1.

to capture subtle patterns in chest X-rays, by recalibrating the importance of feature channels and thus improving diagnostic accuracy.

Interestingly, the improvement brought by SE blocks was less pronounced at extreme values of α. In particular, at $\alpha = 0.25$, the SE-enhanced model had a small decrease in precision compared to the baseline. This is likely because, at this smaller size, the model is already highly efficient, making the channel attention mechanism have less impact.

C. Analysis of Lightweight Configurations

One particularly surprising result was the strong performance of the models with $\alpha = 0.25$, both in their baseline and SE-enhanced forms. Despite having a drastically reduced number of parameters (0.2M for MobileNetV1 and 0.3 SE-MobileNetV1), these models achieved competitive performance with Top-1 accuracies of 96.8% and 96.3%, respectively. This indicates that even highly compact models can maintain excellent predictive accuracy.

The success of the $\alpha = 0.25$ configurations suggests that for specific tasks like lung disease classification from chest X-ray images, it is possible to achieve a favorable balance between computational efficiency and diagnostic accuracy. These lightweight models are particularly well-suited for deployment in resource-constrained environments, where computational resources are limited, yet timely and accurate diagnosis is critical.

D. Visualizing Model Predictions

To further analyze the behavior of our models, we employed CAM and GradCAM to visualize the areas of the chest X-rays that the networks focused on when making their predictions. Fig. 9 shows the activation maps generated for correctly classified COVID-19 cases. To effectively assess the validity of the generated maps, expert support from the medical field would be necessary. However, we can observe that in the examples shown, the model appears to focus on the lung areas, which makes the results seem plausible. Additionally, the heatmaps generated using GradCAM appear more concentrated in specific areas, highlighting a more precise region

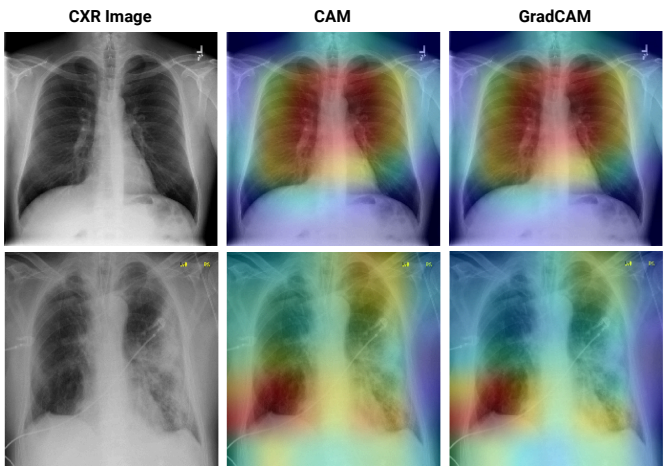

Fig. 9: Class activation visualization using CAM and GradCAM methods. The images in the top row represent a healthy patient ("normal"), while the images in the bottom row correspond to a COVID-19 patient.

of interest. This is crucial in medical applications, where understanding the modelâ€™s reasoning can increase trust in automated diagnostic systems.

VIII. Conclusions

In summary, the experiments demonstrate the effectiveness of lightweight convolutional architectures in the task of lung disease classification. Despite the initial expectation that MobileNetV2 would outperform MobileNetV1 due to its more advanced design, the results indicate that MobileNetV1 is better suited for this specific task. Moreover, the integration of SE blocks has significantly enhanced the performance of both architectures, particularly benefiting MobileNetV1. Furthermore, the use of activation visualization techniques proved effective in gaining more insights into the image areas considered by the model for the final classification.

a) Future work: Looking ahead, we believe it would be interesting to further develop the work in several directions.

- It would be useful to explore different configurations of the architectures by choosing where to position the SE blocks more strategically, perhaps using techniques of NAS, and tuning the reduction ratio r.
- Moreover, considering additional lightweight architectures such as ShuffleNet and EfficientNet might assess their applicability in the specific domain of lung disease classification.
- Finally, adopting other soft attention mechanisms, such as feature pyramid networks to integrate information at various scales and CBAM modules combining spatial and channel attention, could offer significant performance improvements.

b) Personal reflections: From a personal perspective, defining the direction in which to develop the research was the most stimulating phase of this work, and at the same

time, it was also the most challenging. The literature review played a fundamental role, providing me with a deep understanding of the task, existing approaches, and possible future developments. Overall, I find this project to be very constructive, both theoretically and practically, having given me the opportunity to delve into a specific topic such as lightweight CNN networks and attention mechanisms, and to apply it in a specific domain.

REFERENCES

[1] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[2] P. Rajpurkar, J. Irvin, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[3] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[4] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[6] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1314–1324, 2019.

[7] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, 2018.

[8] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, 2019.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[11] A. S. C. Asraf, M. K. Islam, K. A. A. M. Haque, and A. K. Akhter, "Covid-19, pneumonia and normal chest x-ray pa dataset," 2021.

[12] R. contributors, "Radiopaedia covid-19 contributions," 2020.

[13] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, and D. Maffitt, "The cancer imaging archive (tcia) - covid-19 cases," 2020.

[14] S. contributors, "Covid-19 database from the italian society of medical radiology (sirm)," 2020.

[15] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, and A. E. T., "Mendeley - covid-19 radiography database," 2020.

[16] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," 2018.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106, 2017.