Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7546488

COMPARATIVE STUDY OF MOBILENETV2, SIMPLE CNN AND VGG19 FOR IMAGE CLASSIFICATION

S. K. Bharadwaj[1], Rahul Jha[2], Jitendra Kumar [3], D. K. Mishra[4], Vikas Shinde[5], V. K. Jadon[6]

[1, 2, 3,4]Department of Engineering Mathematics and Computing
Madhav Institute of Technology & Science, India
[5]Department of Applied Sciences, Anand Engineering College, Agra

Abstract

In the dynamic field of computer vision and image classification, the choice of a deep learning model plays a crucial role in achieving optimal performance across various applications. This study conducts a comparative analysis of three distinct convolution neural network (CNN) architectures: MobileNetV2, VGG19, and a simplified CNN model. The objective is to assess their effectiveness in image classification tasks. MobileNetV2, recognized for its lightweight design, demonstrates notable efficiency in computational resource usage, making it well-suited for deployment on resource-constrained devices. VGG19, characterized by its deep and intricate structure, exhibits a strong ability to capture complex hierarchical features, albeit with increased computational demands. The simplified CNN model, designed to strike a balance between complexity and performance, emerges as a practical alternative in scenarios where a compromise between accuracy and resource efficiency is sought.

Keywords: Image Classification, Convolution Neural Network (CNN), Comparative Analysis Accuracy (CAA), MobileNetV2, VGG19.

1. INTRODUCTION

Convolution Neural Networks (CNNs) have significantly enhanced the performance of image classification tasks by autonomously learning spatial hierarchies of features. This research paper explores the effectiveness of three CNN architectures: MobileNetV2, VGG19 and a simple CNN model in image classification.

MobileNetV2, developed by Google, is an efficient and lightweight model tailored for mobile and embedded vision applications. It utilizes inverted residuals and linear bottlenecks to balance computational efficiency and model accuracy, crucial in resource-limited mobile and embedded scenarios.

In contrast, VGG19, originating from the Visual Graphics Group at Oxford, is a deeper and more complex model renowned for its outstanding performance on the Image Net dataset. Its depth and complexity, coupled with high performance, make it a popular choice for image classification. However, VGG19 comes with significantly higher computational requirements compared to MobileNetV2.

In contrasting these pre-trained models, a simple CNN model, typically comprising a few convolution layers followed by max-pooling and fully connected layers, offers a more straightforward and customizable approach to image classification. While it may not attain the

same accuracy levels as its more intricate counterparts, its simplicity and lower computational requirements render it an appealing option for specific applications.

The literature review includes important contributions to deep learning and convolution neural networks (CNNs). Krzyzewski et al. [1] laid the foundation for their landmark work on Image Net classification using deep CNN, which significantly improved computer vision. Farabate et al. [2] extended this by introducing hierarchical features for visual labeling, which demonstrated the importance of feature learning in complex tasks. Simonyan and Zisserman [3] pushed the boundaries further with very deep convolution networks, which set the benchmark for large-scale images found. Sandler et al. [4] introduced MobileNetV2, which emphasized efficient performance with inverted residuals and linear bottlenecks. Tan and Le [5] redefined model scaling with Efficient Net, and addressed the trade-off between accuracy and computational cost. Notable applications are mechanical fault detection.Jiao et al. [6] image classification, Sharma & Phonsa, [7], Bansal et al.[8] skin lesion classification. Jasil & Ulagamuthalvi [9] some recent developments focused on special areas such as COVID-19 detection. Kaya & Gürsoy, [10] fruit image classification. Gulzar [11] demonstrate the versatility of CNN models and it continues to operate in various industries.
Despite the valuable insights provided by these works on MobileNetV2, VGG19 and simple CNN models for image classification, a comprehensive comparison of these three models in terms of accuracy, computational efficiency and ease of implementation is still lacking which requires research. The purpose is to address.

2. TOOLS AND METHODOLOGY

2.1. CIFAR-10 Dataset:

The CIFAR-10 dataset, developed by the Canadian Institute for Advanced Research, is a widely used compilation of 68,000 32x32 color images sorted into 10 categories: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. Featuring 6,800 images per category, it stands as a crucial asset for training machine learning and computer vision algorithms, frequently serving as a standard for assessing their effectiveness.

Each image within the CIFAR-10 dataset is a 32x32 color image, with RGB values arranged in row-major order. The dataset is partitioned into five training batches and one test batch, each containing 10,000 images. Notably, the test batch consists of 1,000 randomly chosen images per category, while the training batches may exhibit varying distributions across categories.
Despite its popularity, the CIFAR-10 dataset presents challenges due to the small image size, making it challenging for algorithms to recognize subtle patterns. Additionally, the limited number of images per class poses difficulties in distinguishing between similar-looking classes.

However, these challenges render the CIFAR-10 dataset an excellent tool for testing the robustness of machine learning algorithms. Success on CIFAR-10 suggests potential efficacy on larger, more intricate datasets.

In summary, the CIFAR-10 dataset is a valuable asset for machine learning and computer vision enthusiasts. Its compact size, complexity, and widespread use make it an effective

benchmark for assessing the performance of machine learning algorithms, catering to both beginners and seasoned researchers in the field.

2.2. Simple CNN:

In the described Convolution Neural Network (CNN) architecture, a strategic design is employed to enhance feature extraction from input data, particularly in the context of images. The architecture consists of three distinct blocks, each comprising convolution layers and max pooling layers. This design facilitates effective hierarchical representation learning, allowing the network to discern intricate patterns within the data.

The operations within each block follow a specific sequence:

2.2.1. Convolution Layer: This layer employs filters for convolution over the input data, capturing spatial hierarchies and extracting meaningful features. The convolution operation allows the network to identify local patterns and acquire representations of features.

2.2.2. Max Pooling Layer: After the convolution process, the subsequent step involves applying a max pooling layer to decrease the spatial dimensions of the feature maps. Max pooling entails selecting the maximum value from a group of neighboring values, effectively down sampling the information while retaining the most significant features. This helps manage the computational complexity of the network and improves its robustness.

After these blocks, the feature maps undergo flattening, transforming the spatial information from the convolution and pooling layers to a one-dimensional vector. This flattening operation is crucial as it prepares the data for input into densely connected layers.

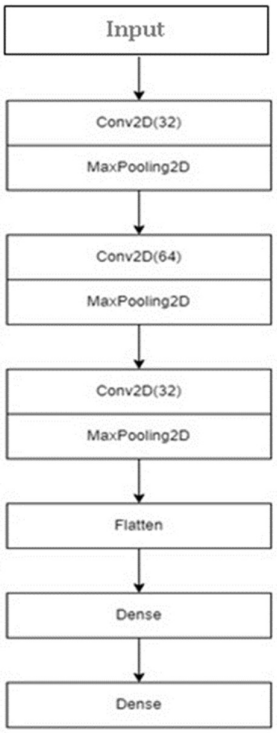

Fig 1: Block Diagram of simple CNN

2.3. MobileNetV2:

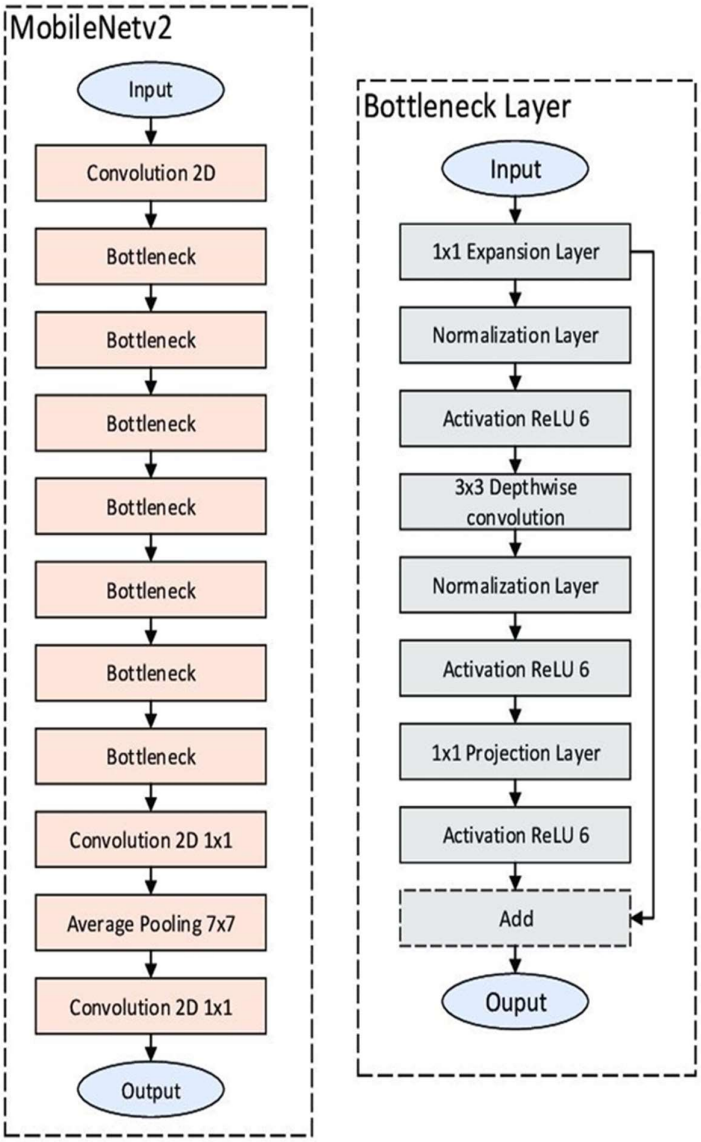

Fig 2: Block Diagram of MobileNetV2

An evolution of its predecessor, Mobile Net is recognized for its exceptional efficiency in resource-constrained environments like mobile devices and edge computing platforms. Key to its innovation is the integration of depth wise separable convolutions significantly reducing parameters and computations compared to traditional convolution layers.

The incorporation of MobileNetV2 as the foundation layer combined with global average pooling and a dense layer which signifies a thoughtful design strategy that addresses both computational efficiency and nuanced feature representation.

Global average pooling computes the average value of each feature map across its spatial dimensions, reducing spatial dimensions to a single value per channel.

The dense layer leverages features extracted by MobileNetV2 and condensed by global average pooling to make nuanced predictions. Its configuration corresponds to the number of output classes with an activation function often soft-max, is converting raw output into probability distributions.
The overall harmonization of these components signifies a deliberate effort to balance computational efficiency model compactness and classification accuracy.

2.4. VGG19:
VGG19, a derivative of the Visual Geometry Group (VGG) architecture which is renowned for its simplicity and efficacy. Comprising 19 layers, it is predominantly constructed with 3x3 convolution filters and interspersed with max-pooling layers. The iterative arrangement of these layers enables VGG19 to effectively capture intricate hierarchical features within input data, establishing it as a formidable option for tasks related to images.
As the foundation layer, VGG19 serves as a feature extractor, hierarchically learning and representing complex patterns and structures within the input data.

After the foundation layer of VGG19, a global average pooling layer is incorporated. Global average pooling functions is calculating by the average value for each feature map across its complete spatial dimensions. Additionally, the reduction in spatial dimensions are achieved by global average pooling facilitates computational efficiency in subsequent layers.

After the pooling layer of VGG19 a dense layer, functioning as the ultimate phase for classification. This fully connected layer receives the compact features derived from the global average pooling layer and associates them with the designated output classes or predictions.

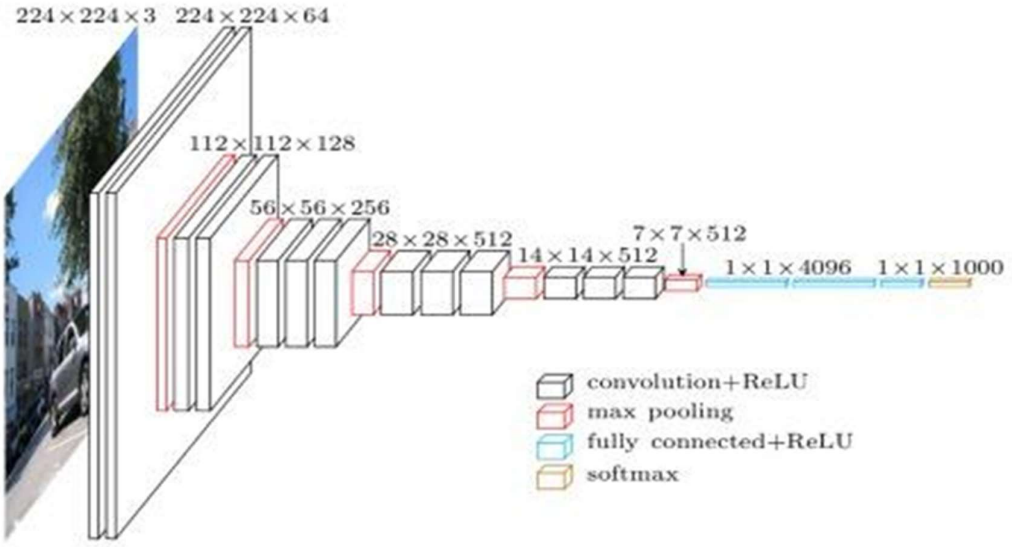

Fig 3: VGG-19 Architecture [12]

3. RESULTS

The assessment and training of MobileNetV2, VGG19 and simplified CNN models were conducted using the CIFAR-10 dataset, which contains 68,000images, 32x32 color images categorized into ten categories which is a widely recognized and frequently employed benchmark dataset. Tailored for tasks related to image classification to evaluate and comparison of three different CNN architecture. The central emphasis of this analysis rests on the accuracy and confusion matrices of each model.

Models	Models param	data size	param/data
CNN	68,000	10,000	68.0
MobileNetV2	2270794	10,000	227.0794
VGG19	20034644	10,000	2003.4644

Table 1: Comparison of Network Variables of Various CNN's

Following graph presents the accuracy results for the three models on the CIFAR-10 dataset.

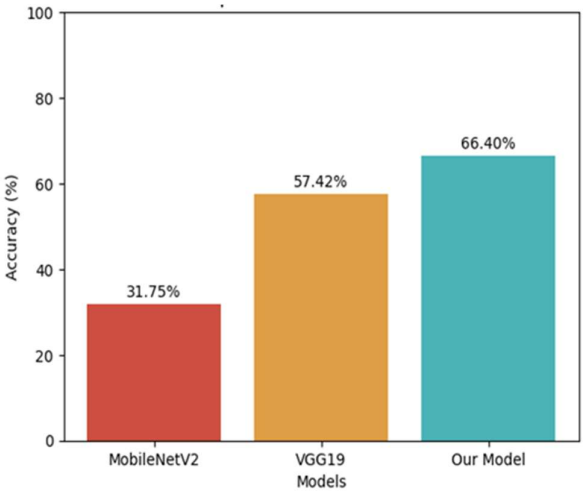

Fig 4: Comparison of Model Accuracies

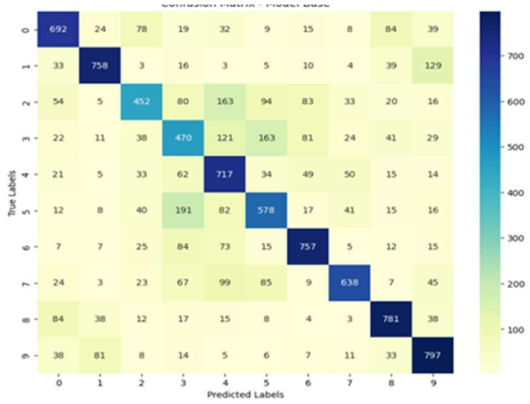

Fig 5: Confusion matrix of CNN

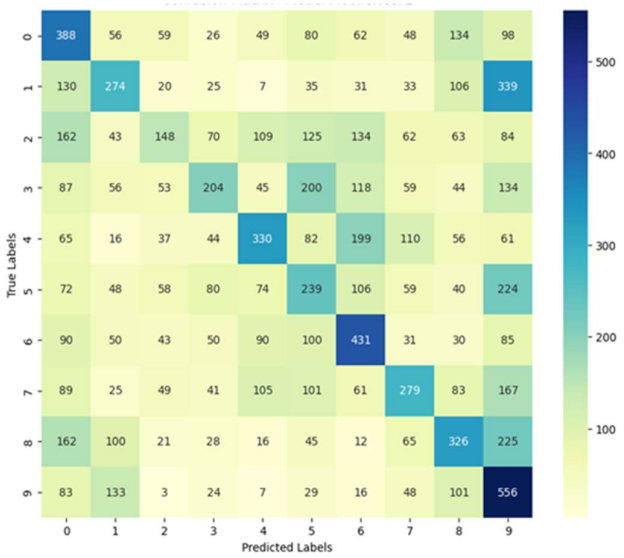

Fig 6: Confusion matrix of MobileNetV2 Model

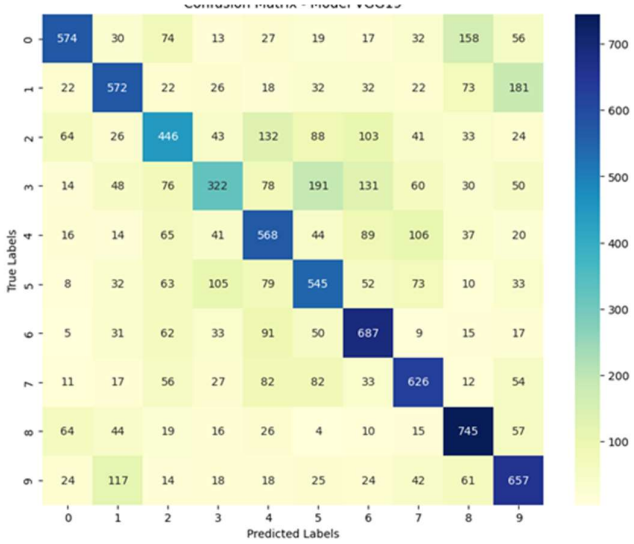

Fig 7: Confusion matrix of VGG19 Model

The findings are visually depicted in the subsequent graphs, offering a thorough examination of the model's performance across the entirety of the training process. The initial set of graphs presents summarized loss in both normal and log scales, providing insights into the overall convergence trend. Following this, a more detailed analysis follows with additional graphs showing individual training and validation losses. These visualizations contribute to a nuanced comprehension of the model's behavior, illuminating its training dynamics and generalization capabilities.

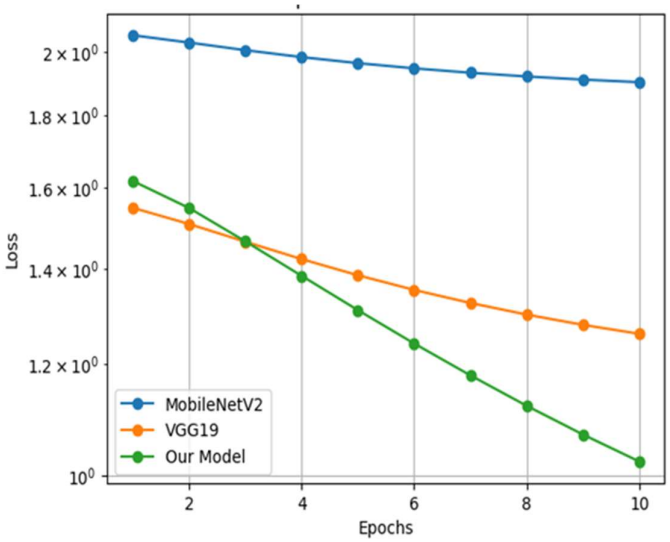

Fig 8: Comparison of Model Losses

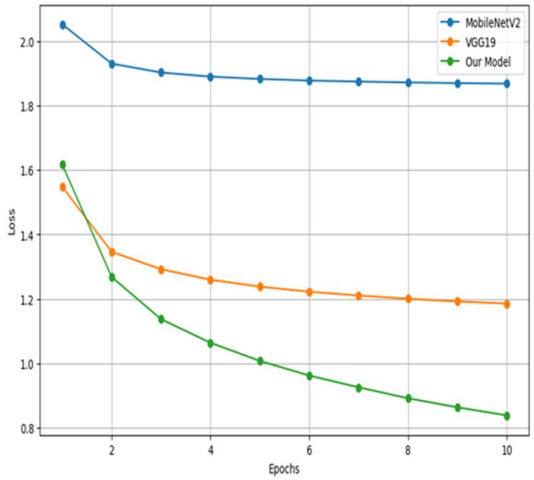

Fig 9: Comparison of Model Losses

The loss comparison of individual function is also illustrated by following figures

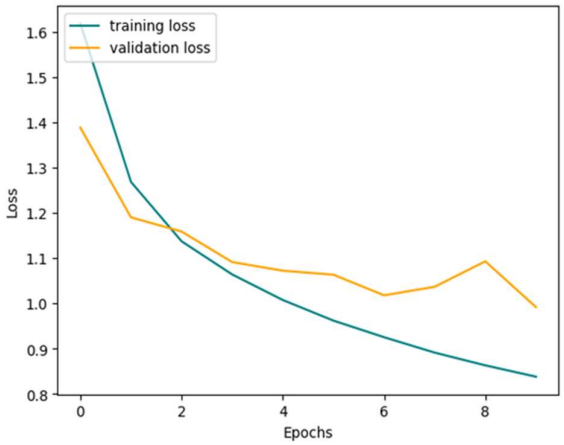

Fig10: CNN model Loss

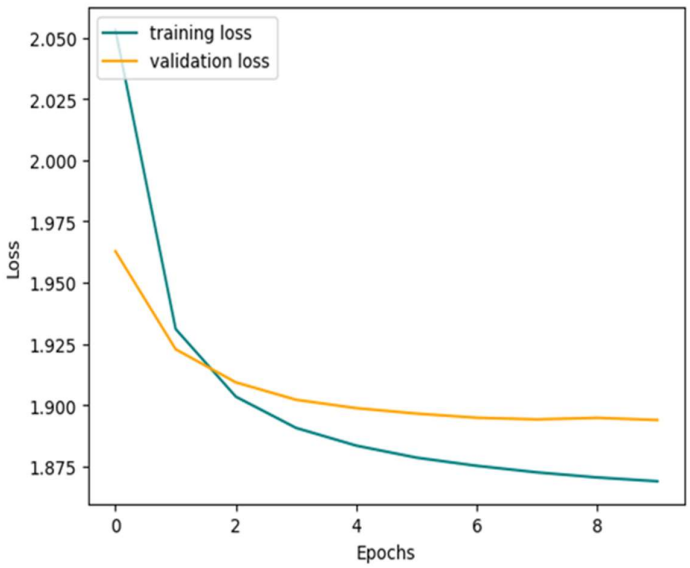

Fig 11: Mobile NetV2 Model Loss

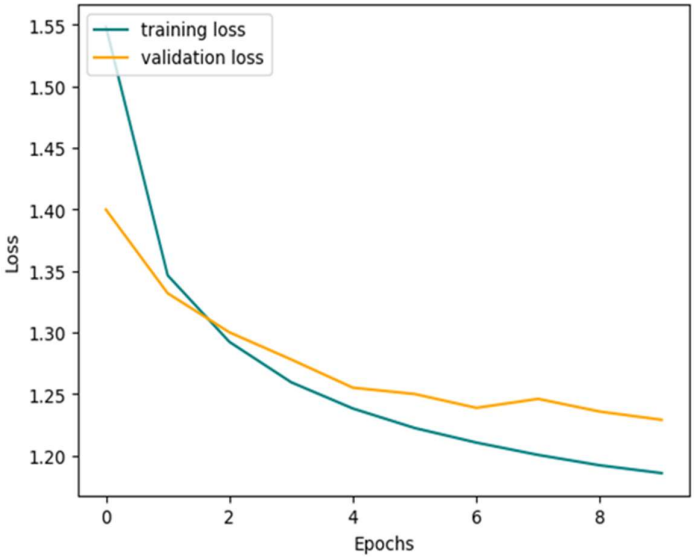

Fig 12:VGG19 Model Loss

The AUC-ROC (Area under the Receiver Operating Characteristic) score stands as a widely employed metric for assessing the performance of models in binary classification scenarios. It gauges a model's ability to differentiate between positive and negative classes under varying probability thresholds.

In the subsequent section, the table provided enumerates the AUC-ROC scores of the models, offering a comprehensive assessment of their proficiency in discriminating between positive and negative classes.

Models	AUC-ROC Score

CNN	0.9421106
MobileNetV2	0.7725232888888888
VGG19	0.9116561666666667

Table 2: Comparison of Network Variables of Various CNNs

The study findings reveal that among three models assessed, a simple CNN model attains the highest accuracy. This implies that a streamlined and adaptable architecture can be highly effective for image classification tasks. The added advantage of the simple CNN model lies in its lower computational cost, emphasizing its practical utility. Although, the MobileNetV2 and VGG19 models exhibit comparatively lower accuracies in this study, it's worth noting that their pre-trained weights and transfer learning capabilities may confer advantages for particular image classification tasks.

4. CONCLUSION

In this article, we evaluate and compare the effectiveness of three different models convolution neural network (CNN) architectures: MobilenetV2, VGG19 and a simple CNN model. Then we observed that the simple CNN model has much better accuracy with result of mobilenetV2 and VGG19 and the computational expense of simple CNN is much lower than mobilenetV2 and VGG19.

REFERENCES

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

[2] Farabet, Clement, Camille Couprie, Laurent Najman, and Yann LeCun. "Learning hierarchical features for scene labeling." IEEE transactions on pattern analysis and machine intelligence 35, no. 8 (2012): 1915-1929.

[3] Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556.

[4] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, (pp. 4510-4520).

[5] Tan, M., & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, 2019, (pp. 6105-6114). PMLR.

[6] Jiao, J., Zhao, M., Lin, J., & Liang, K. . A comprehensive review on convolutional neural network in machine fault diagnosis. Neurocomputing, 2020, 417, 36-63.

[7] Sharma, A., & Phonsa, G. Image classification using CNN. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC), 2021.

[8] Bansal, M., Kumar, M., Sachdeva, M., & Mittal, A. . Transfer learning for image classification using VGG19: Caltech-101 image data set. Journal of ambient intelligence and humanized computing, 2021, 1-12.

[9] Jasil, S. G., & Ulagamuthalvi, V. . Deep learning architecture using transfer learning for classification of skin lesions. Journal of Ambient Intelligence and Humanized Computing,2021, 1-8.
[10] Kaya, Y., & Gürsoy, E. A MobileNet-based CNN model with a novel fine-tuning mechanism for COVID-19 infection detection. Soft Computing, , 2023, 27(9), 5521-5535.
[11] Gulzar, Y. . Fruit image classification model based on MobileNetV2 with deep transfer learning technique. Sustainability, 2023, 15(3), 1906.
[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.