Exploring Prominent Convolutional Neural Network Frameworks to Identify COVID-19 Deceases by Using Medical Images

Yallapu Srinivas[1,2(✉)] and M. Aravind Kumar[3]

[1] Bharatiya Engineering Science and Technology Innovation University, Gownivaripalli, Gorantla, Andhra Pradesh 515231, India
Yallapu.srinivas@gmail.com
[2] Department of ECE, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh 534202, India
[3] West Godavari Institute of Science and Engineering, Prakashraopalem, East Godavari District, Andhra Pradesh, India

Abstract. Computer vision and image classification have been used significantly in the clinical field, due to the availability and implementation of various Convolutional Neural Networks (CNNs) over the past decade. Hence, we present an analysis report on several prominent CNN architectures such as AlexNet, VGGNet, Inception (GoogLeNet), ResNet, EfficientNet, RegNet, ViT (Vision Transformer), and Swin Transformer by exploring their historical context, architectural details, and key innovations. Finally, we aim to assist researchers and practitioners in choosing the most appropriate architecture by comparing the accuracy, trainable parameters, and computational requirements of aforementioned architectures to identify COVID-19 from chest X-ray images for further clinical process/specific research.

Keywords: Convolutional Neural Network · AlexNet · VGGNet · ResNet · Inception · Deep learning · EfficientNet · RegNet · ViT (Vision Transformer) · Swin Transformer

1 Introduction

The entire world was affected by coronavirus in 2019. According to the World Health Organisation statistics total of 775 billion people and 7 billion people were dead. Detection of COVID-19 through analysis of chest X-ray images is a significant application of CNNs. CNNs are particularly well-suited for image classification tasks like identifying patterns in medical images. In the context of COVID-19, CNNs can be trained on datasets of chest X-ray images to learn features that distinguish between X-rays of patients with COVID-19 and those without. By training CNNs on a dataset of chest X-ray images showing signs of COVID-19, researchers and medical professionals can

P. Pareek et al. (Eds.): IC4S 2024, LNICST 597, pp. 184–203, 2025.
https://doi.org/10.1007/978-3-031-77075-3_16

develop models that help automate the detection process. These models can assist radiologists in quickly and accurately identifying cases of COVID-19, potentially speeding up diagnosis and improving patient outcomes. Furthermore, CNNs can be used not only for COVID-19 detection but also for tasks like disease progression monitoring, severity assessment, and treatment evaluation. Convolutional Neural Networks (CNNs) have significantly advanced image classification, with several seminal architectures emerging over the years.

This survey delves into the details of four prominent CNNs - AlexNet, VGGNet, ResNet, Inception EfficientNet, RegNet, ViT (Vision Transformer), and Swin Transformer. By understanding their historical significance, motivations, structural components, and innovative features, we aim to provide insights into the progress of deep learning in image classification. Utilizing transfer learning to extract knowledge from diverse domains is highly advantageous. Transfer learning involves the reuse of a pre-trained model's knowledge for various tasks, such as classification, regression, and clustering. VGG-16 is a Deep Convolutional Neural Network, to perform image classification [1, 2].

An enhanced U-Net model incorporating VGG-16 for the precise segmentation of Brain MRI images, specifically to identify regions of interest, such as tumor cells. Our comparative analysis is according to the TCGA-LGG dataset the accomplishment of common state-of-the-art CNN-based approaches in this field [3]. Introducing a novel CNN design called P_VggNet, which combines two key components: P_Net and VggNet-16, the latter being a 16-layer variant of the VggNet. The design of P_Net and its integration into the P_VggNet structure were carefully engineered [4]. The identification of severity levels was based on computed ratios derived from segmented images. These parameters included the mean and standard deviation of pixel intensities, as well as the mean values of hue and saturation. Deviation clustering was also computed. These derived features were then utilized as input for a classifier to perform diabetic retinopathy (DR) classification. Specifically, a VGG-19 deep neural network was trained and tested using these derived feature sets [5].

The survey provides a concise overview of the progress made in the field of Deep Learning (DL), commencing with the inception of Deep Neural Networks (DNN). It subsequently explores advancements in CNN, Recurrent Neural Networks (RNN) with a focus on Long short-term memory (LSTM) and Gated Recurrent Units (GRU), Auto-Encoders (AE), Deep Belief Networks (DBN), Generative Adversarial Networks (GAN), and Deep Reinforcement Learning (DRL) [6]. The performance was assessed across four experimental schemes, each varying the degree of knowledge acquired from the pre-trained model in the experiments [7]. To expedite convergence, the Convolutional blocks incorporate the Exponential Linear Unit (ELU) activation function. To mitigate overfitting, Dropout is strategically implemented across various layers within the network [8].

A highly accurate classification of pressure and diabetic wound images was achieved. Six distinct variations of the new AlexNet architectures were developed, each featuring varying implementations of Convolution, Pooling, and Rectified Linear Activation (ReLU) layers. The performance of these proposed models was thoroughly examined using both Softmax and SVM classifiers separately [9]. The swift integration of deep

learning in image registration applications in recent years underscores the need for a thorough overview and prospects. Emphasis will be placed on various research domains, along with the identification of challenges encountered by practitioners [10–14]. This discussion so far finalized various Convolutional Neural Network Architectures. The performance of CNN depends upon the following metrics.

- Accuracy: The proportion of correctly classified images.
- Precision: The proportion of correctly classified positive instances among all instances classified as positive.
- Recall (Sensitivity): The proportion of correctly classified positive instances among all actual positive instances.
- F1 Score: The harmonic mean of precision and recall.
- Specificity: The proportion of correctly classified negative instances among all actual negative instances.

1.1 AlexNet

AlexNet is a convolutional neural network (CNN) architecture that gained significant attention and marked a breakthrough in the field of computer vision, particularly in image classification. The Fig. 1 shows the block diagram of the Alexnet.

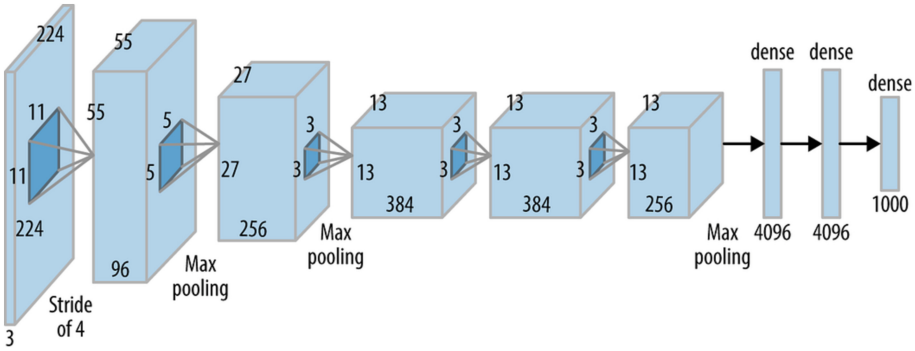

Fig. 1. Alexnet Block Diagram.

The network consists of 62 million trainable variables. Data augmentation is employed to mitigate overfitting, involving techniques such as mirroring and cropping images to enhance diversity within the dataset used for training. The network incorporates max-pooling with overlap layers following the initial, second, and fifth convolution (CONV) layers. Max-pooling with overlap layers, characterized by strides smaller than the window size, uses a 3×3 max-pool layer with a step size of 2, resulting in overlapping receptive fields. This overlap contributed to a 0.4% improvement in top-1 errors and a 0.3% enhancement in top-5 errors. Before the introduction of AlexNet, the prevailing activation functions included sigmoid and tanh. These functions, due to their saturation characteristics, encountered the Vanishing Gradient (VG) problem, posing challenges for effective network training.

AlexNet addresses this issue by employing the Rectified Linear Unit (ReLU) node activation, which does not experience challenges related to the VG issue. The preliminary investigation showed that a network employing ReLU achieved a 25% error rate about six times faster than a comparable network using tanh activation.While Rectified Linear Unit, commonly known as ReLU is effective in mitigating the issue of gradient vanishing, its unbounded nature may lead to excessively high learned variables. To address this issue, AlexNet implemented Local Response Normalization (LRN), a technique that normalizes a pixel's neighborhood, enhancing the activated neuron while simultaneously suppressing its surrounding neurons. In addition, AlexNet tackles the overfitting challenge through the incorporation of dropout layers. During training, connections are randomly dropped with a chance of $p = 0.5$, dropout is employed. Although dropout prevents overfitting by aiding the network in avoiding undesirable local troughs, it concurrently increases the number of iterations needed for culmination.

The authors achieved an accuracy of around 80–90% for classifying skin cancer using dermoscopy images application of AlexNet [15]. The 95% of accuracy by the automated Alzheimer's disease classification can be helpful as an assisting tool for medical personnel to diagnose the stage of Alzheimer's disease [16].

1.2 VGGNet

VGG (Visual Geometry Group) Net is a convolutional neural network (CNN) architecture that was proposed by the Visual Geometry Group at the University of Oxford. While it did not win the competition, it demonstrated the effectiveness of using deep convolutional neural networks for image classification tasks. A comparison of different versions of VGGNet up to the latest versions is given below.

Table 1. Comparison of the VGG architectures.

VGGNet Version	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	FLOPs (Billion)
VGG11	2014	69.00%	88.50%	132	7
VGG13	2014	69.40%	88.90%	133	11.2
VGG16	2014	71.50%	89.80%	138	15.3
VGG19	2014	71.60%	89.80%	143	19.4

VGGNet is known for its simplicity and uniform architecture, with all convolutional layers using 3×3 filters and all max-pooling layers using 2×2 filters with stride 2. As the version number increases, the network depth increases, leading to better performance but also increased computational cost and memory requirements. VGGNet achieved competitive performance and was evaluated on the ImageNet dataset at the time of its introduction and served as a basis for deeper architectures.

The Fig. 2 shows the block diagram of VGG16. It comprises a total of 138 million parameters. It's crucial to highlight that all convolutional kernels have a size of $3 \times$

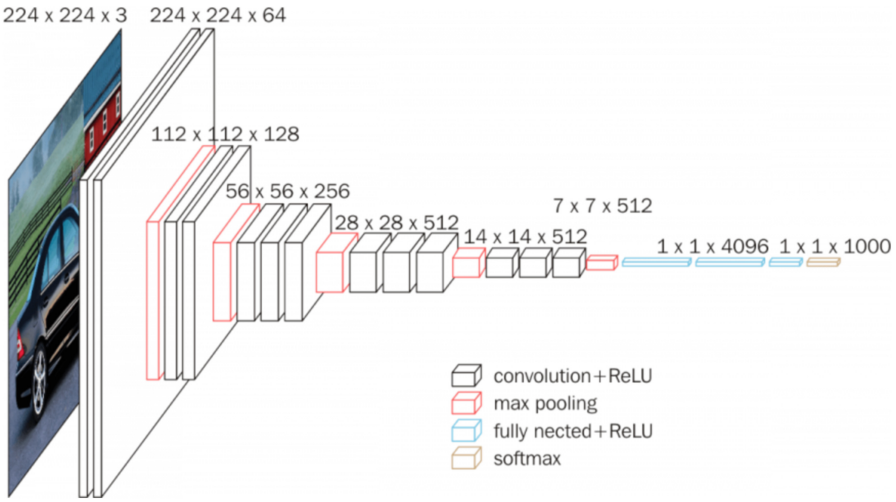

Fig. 2. Block diagram of VGG16.

3, and the max-pooling kernels measure 2 × 2 with a stride of two. Take into account the provided example: suppose we possess an input layer with dimensions 5 × 5 × 1. Applying a convolutional layer employing a 5 × 5 kernel and A stride of one results in an output feature map with a specific size of 1 × 1. Alternatively, achieving Two consecutive operations enables the generation of the same output feature map 3x3 convolutional layers demonstrating a stride of 1, as depicted below in Fig. 3.

For a 3 × 3 convolutional layer with 2 channels, the total number of variables is 18 (3 × 3 × 2). Conversely, employing a single 5 × 5 convolutional layer necessitates 25 variables, indicating a reduction of 28%. On the other hand, achieving the impact of a 7 × 7 (or 11 × 11) convolutional layer may be accomplished by incorporating 3 (or 5) consecutive 3 × 3 convolutional layers with a step of 1. This strategic approach results in a noteworthy reduction of trainable variables by 44.9% (for three layers) or 62.8% (for five layers). A diminished number of trainable variables not only facilitates faster learning but also enhances resilience against overfitting. VGGNet, with its deep architecture, has been applied to various medical image analysis tasks, including disease identification.

- Cancer: VGGNet has been used to identify various types of cancer, including breast cancer, lung cancer, and skin cancer, with reported accuracies ranging from 80% to 95% or higher depending on the dataset and specific implementation [17].
- Neurological Disorders: VGGNet has been applied to the identification of neurological disorders such as Alzheimer's disease and Parkinson's disease, with reported accuracies in the range of 85% to 95% [18].
- Ophthalmic Diseases: VGGNet has been used for the identification of diabetic retinopathy, glaucoma, and other ophthalmic diseases, achieving accuracies of around 94% to 96%. The authors proposed an improved model that was trained on the iChallenge-GONdataset, trained the ResNet, AlexNet, and VGGnet work models on the three data sets were compared to evaluate the proposed model [19, 20].

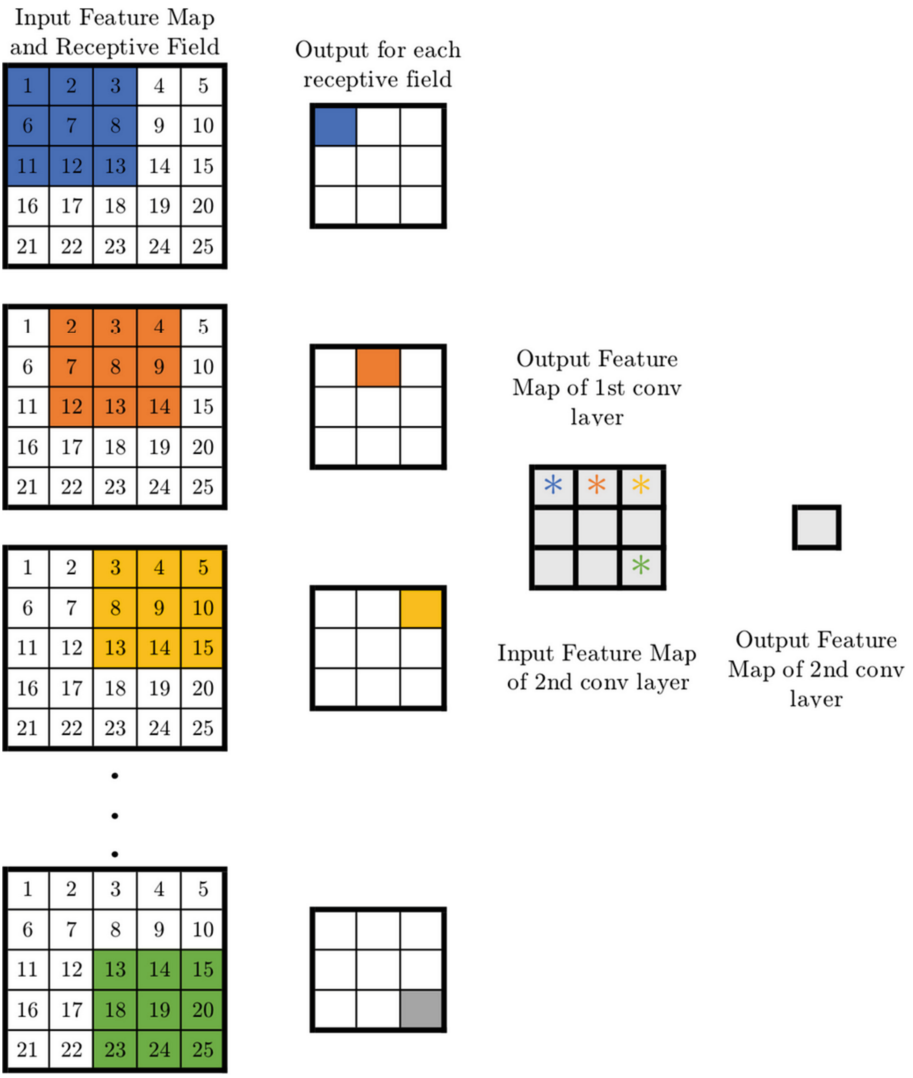

Fig. 3. The output feature map is implemented by using two consecutive 3 × 3 convolutional layers utilizing a stride of 1.

1.3 Inception

Inception, also known as GoogLeNet, is a deep learning construction designed for image recognition and classification tasks. It was introduced by researchers at Google, led by Christian Szegedy, in their 2014 paper "Going Deeper with Convolutions." Inception stands out for its innovative use of inception modules, which employ filters of different sizes within the same layer to capture information at multiple scales. These details are shown in the Fig. 4.

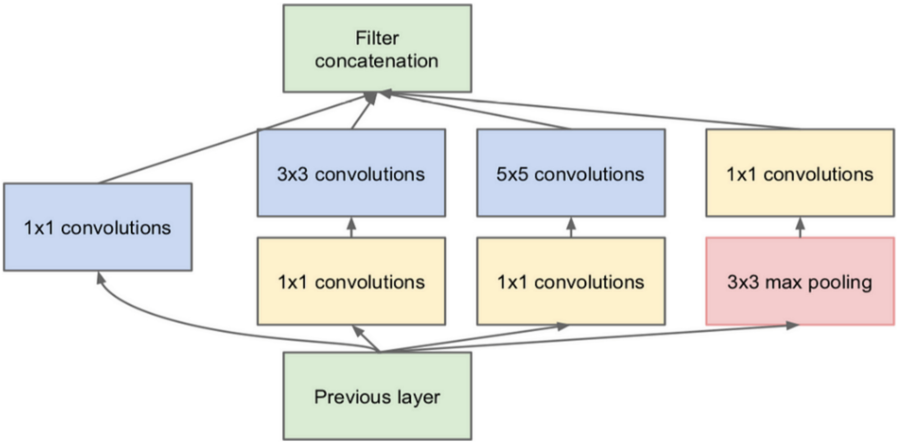

Fig. 4. Inception Module.

The module of inception is composed of four concurrent operations.

- 1×1 convolution layer,
- 3×3 convolution layer,
- 5×5 convolution,
- Max pooling.

The 1×1 convolutional hunks, depicted in yellow, serve the purpose of depth reduction. The outcomes emerging from the four concurrent operations are subsequently merged in terms of depth to create the block for concatenating filters, highlighted in green. Various versions of Inception exist, with the simplest being the GoogLeNet, It is a CNN architecture known for its innovative use of "Inception modules" to improve efficiency and performance. Here's an analysis of its performance metrics and efficiency for classifying COVID-19 medical images, Performance Efficiency: Inception models are designed to be efficient while maintaining high performance. They achieve this through their use of multiple parallel convolutional paths within each Inception module, allowing the network to capture features at different scales efficiently. Inception models, especially later versions like Inception-v3 or Inception-v4, are known for their strong performance on image classification tasks. Comparison of different versions of Inception (GoogLeNet) up to the latest versions:

- The Inception architecture, also known as GoogLeNet, is characterized by its inception modules, which allow for the parallel operation of filters of different sizes within the same convolutional layer.
- Inception-v1 was the original architecture introduced in 2014, and subsequent versions (v2, v3, v4, and Inception-ResNet-v2) improved upon it by adding deeper architectures and incorporating residual connections.

1.4 ResNet

ResNet, short for Residual Networks, is a deep learning architecture that was introduced to address the challenges of training very deep neural networks [21–26]. Consider a

Table 2. Comparison of the Inception architectures.

Inception Version	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	FLOPs (Billion)
Inception-v1	2014	69.80%	89.20%	6.7	1.5
Inception-v2	2015	73.90%	91.80%	11.2	2
Inception-v3	2015	77.90%	93.70%	27.2	5.7
Inception-v4	2016	80.20%	95.20%	42.7	12
Inception-ResNet-v2	2016	80.40%	95.30%	55.8	12

scenario where we aim to train A neural network(net) with fully connected layers in a multi-layer perceptron architecture on a dataset when the input is equivalent to the output. The seemingly straightforward result involves setting all weights to one and biases to zero for hidden layers. However, during backpropagation training, this simplistic approach leads to a complex mapping with a wide range of weight and bias values.

Similarly, expanding layers in an already established neural network poses challenges. If We possess a network, denoted as f(x), achieving a dataset with an n% level of accuracy, incorporating additional layers into form g(f(x)) might ideally maintain or enhance accuracy. Regrettably, experiments reveal that accuracy often decreases when additional layers are introduced. These issues stem from the vanishing gradient problem, especially evident as convolutional neural networks (CNNs) increase in depth. Deeper networks result in backpropagated derivatives becoming nearly negligible in value for initial layers. ResNet addresses this challenge through the introduction of two categories of 'shortcut connections': Identity-based shortcuts and Projection-based shortcuts. Before going to the ResNet description we have to see the architectural details of the Residual block (see Fig. 5).

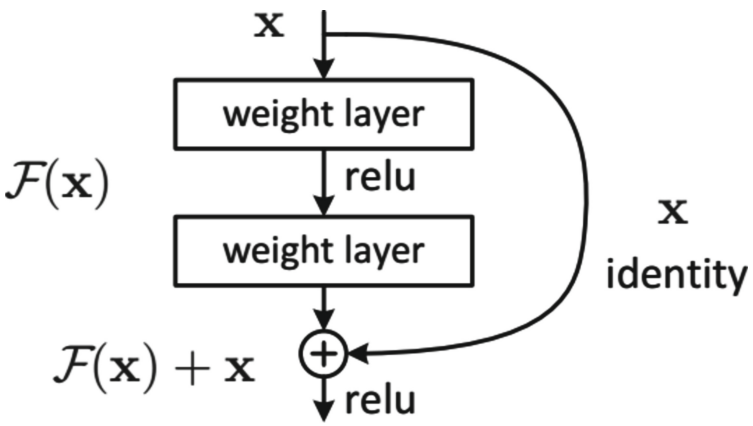

Fig. 5. Residual Block.

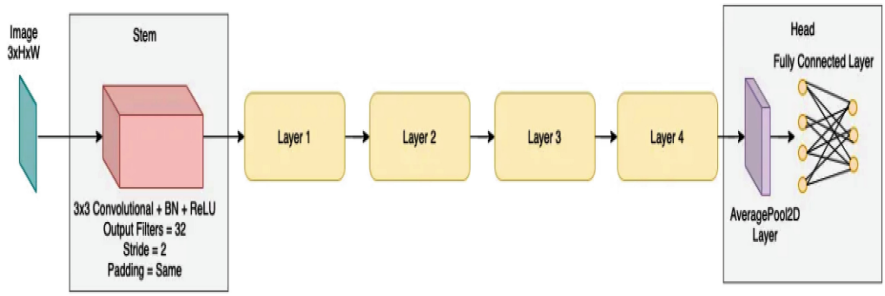

Fig. 6. Architecture of ResNet [27] (Colour figure online).

Figure 6 depicts ResNet architecture and head block, layer block, and stem block.

Stem Block: This block consists of a convolutional layer followed by batch normalization and ReLU activation, with a stride of 2 and a filter size of 3. The number of output filters is either 32 or 64, depending on the specific requirements (referred to as the Red Block in Fig. 6).

Layer Block: The Layer Block consists of chains of residual blocks (depicted in blue in Fig. 7). Figure 7 illustrates the block diagram of these chains within a Layer Block. Let's define the number of blocks in a layer as the depth, denoted as d. The number of channels in each layer remains constant throughout a specific Layer Block and is denoted by the width, denoted as w, in the referenced paper. Each layer within the block takes an input feature map of size W1 $\times$ R $\times$ R (where the first block of every layer converts channels from W1 to W2 as shown in Fig. 2, and then all subsequent blocks output the same number of channels, denoted as W2), and produces an output feature map of size W2 $\times$ R/2 $\times$ R/2, as shown in Fig. 7. Rather than acquiring knowledge of the mapping directly from x to F(x), the net is designed to learn and find the correlation between x and F(x) + G(x). Upon considering the input x and its corresponding output F(x) share identical dimensions, the function G(x) becomes a unique function, and the corresponding shortcut connection is termed an Identity connection.

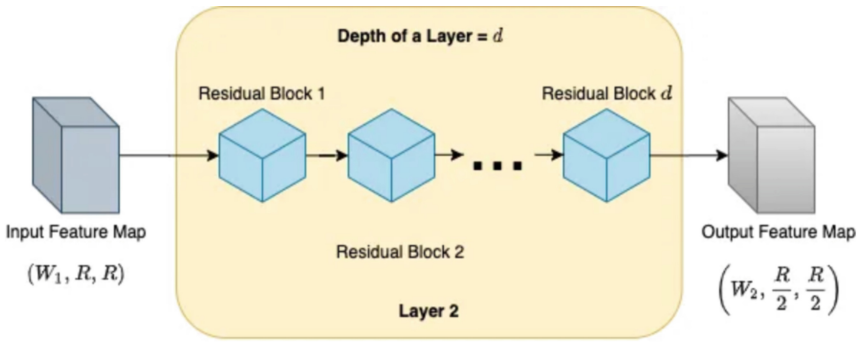

Fig. 7. Sequences of residual blocks within a Layer Block (indicated by the yellow color block in Fig. 6) [27] (Colour figure online).

To achieve an alike charting, the parameters within the intermediate layers are zeroed out during working out, as it is more feasible than pushing them to one. In situations where the extents of F(x) diverge from x, often owing to a pace distance greater than one in the convolution layers in amid, Prediction linking is employed instead of uniqueness linking. The function G(x) is then responsible for adjusting the extent of the excitation x to match those of the response F(x). There are 2 types of mappings in this context.

Non-trainable Mapping (Padding): To ensure matching dimensions with F(x), zeros are simply padded to the input x.

Trainable Mapping (Convolution Layer): A 1×1 Convolutional layer is employed to establish the charting from x to G(x). Examining the table provided reveals a consistent pattern throughout the net where three-dimensional dimensions are either maintained or shared, and the penetration is either retained or doubled. The product of Width and Depth after each convolutional layer remains constant, specifically at 3584. 1×1 convolutional layers are strategically used to halve the three-dimensional and make as twice the depth, achieved through a pace length of 2 and multiples of such filters. The count of 1×1 convolutional layers aligns with the complexity of F(x). The Table 3 shows the comparison of the Inception architectures in terms of accuracy, parameters, and FLOPs (Table 1 and Table 2).

Table 3. Comparison of the Inception architectures.

ResNet Version	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	FLOPs (Billion)
WRN-18 (ResNet-18)	2015	25.58	8.06	45.6	6.70
ResNet-34	2015	25.30	7.98	21.8	3.6
ResNet-50	2015	24.81	7.78	26.6	4.14
ResNet-101	2015	23.6	7.1	44.5	7.51
ResNet-152	2015	23.1	6.97	60.2	11.3
ResNet-200	2016	22.6	6.31	64.9	8.6

- ResNet-18 and ResNet-34 are smaller models suitable for simpler tasks or where computational resources are limited.
- ResNet-50 is a popular choice due to its balance between accuracy and computational cost.
- Deeper models such as ResNet-101, ResNet-152, and ResNet-200 can capture more intricate features but demand greater computational resources.
- ResNet-200: Achieves a top-1 accuracy of 78.6% and a top-5 accuracy of 94.4%. It has 64.9 million parameters and requires approximately 18.6 billion FLOPs for inference.

1.5 EfficientNet

It is a family of CNN architectures that are known for their efficiency in terms of computational resources while maintaining high performance. While EfficientNet was not specifically designed for medical image classification, it can be adapted for this task, including COVID-19 medical images [28, 29]. Different slicings of EfficientNet as shown in Fig. 8.

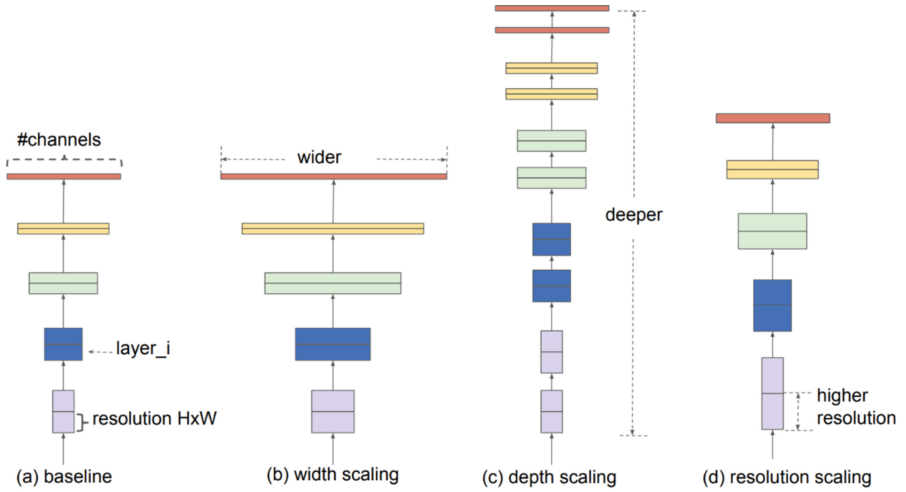

Fig. 8. Model scaling involves different approaches: (a) starting with a baseline network; and (b)-(d) applying conventional scaling, which increases one dimension of the network, such as width, depth, or resolution [28].

Table 4 shows the comparison of different versions of Efficient Net in terms of accuracy, parameters and FLOPs.

- EfficientNet-B0 is the base model with relatively fewer parameters and computations.
- As the version number increases, the models become larger and more accurate, but also more computationally expensive.
- EfficientNet-B7 is the largest and most accurate variant, but it requires significantly more resources for training and inference compared to B0.

1.6 RegNet

RegNet is the Most Flexible Network Architecture For Computer Vision. A model design that scales for high-efficiency or high-accuracy. Traditionally, convolutional neural network architectures have been designed and optimized for one specific purpose. For example, the ResNet model family was optimized for the highest accuracy on ImageNet at the time of its initial release. MobileNets, as the name suggests, are optimized to run on mobile devices. Lastly, EfficientNet was designed to be highly efficient for visual recognition tasks. Authors "Designing Network Design Spaces", [30] decided to set a

Table 4. Comparison of the EfficientNet architectures.

EfficientNet Version	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	FLOPs (Billion)
EfficientNet-B0	2019	84.40%	98.10%	5.3	0.39
EfficientNet-B1	2019	85.50%	98.30%	7.8	0.71
EfficientNet-B2	2019	86.50%	98.60%	9.2	1
EfficientNet-B3	2019	87.30%	98.70%	12.2	1.8
EfficientNet-B4	2019	88.00%	98.90%	19.4	4.2
EfficientNet-B5	2019	88.50%	99.00%	30	10
EfficientNet-B6	2019	88.70%	99.10%	43	19
EfficientNet-B7	2019	88.80%	99.10%	66	37

very unusual but highly interesting goal: They set out to explore and design a highly flexible network architecture. Figure 9 shows the block diagram of the RegNet module and the bottleneck RegNet.

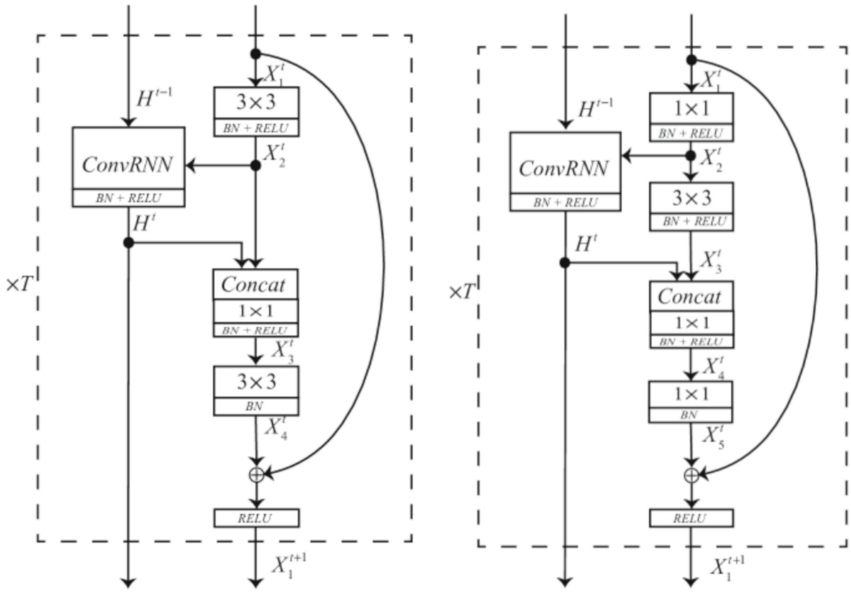

Fig. 9. The RegNet module and the bottleneck RegNet block. T denotes the number of building blocks as well as the total time steps of ConvRNN [31].

One that can be adapted to be highly efficient or run on mobile devices, but also be highly accurate when adapted for the best classification performance. Visualization of network design spaces is constantly optimized to arrive at a smaller design space

with the best models. The approach they took was also very non-traditional: Instead of hand-crafting the model architecture, they set up what they call Network Design Spaces. RegNet is not an architecture, but a network design space. Comparison of different versions of RegNet up to the latest versions as shown it Table 5.

Table 5. Comparison of the RegNet versions.

RegNet Version	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	FLOPs (Billion)
RegNetX-200MF	2020	77.00%	–	17.7	2
RegNetX-400MF	2020	78.60%	–	29.4	4
RegNetY-400MF	2020	79.30%	–	29.9	4
RegNetY-800MF	2020	79.80%	–	49.4	8.1
RegNetY-1.6GF	2020	80.40%	–	88.5	16
RegNetY-3.2GF	2020	81.10%	–	159.3	31.2
RegNetY-4GF	2020	81.50%	–	176.2	39.2
RegNetY-6.4GF	2020	82.00%	–	303.4	78.4
RegNetY-8GF	2020	82.30%	–	400.9	105.4
RegNetY-12GF	2020	82.60%	–	632.3	189.9
RegNetY-16GF	2020	82.80%	–	783.1	302.8

- RegNet is a family of neural network architectures designed for efficient and effective training on a wide range of tasks.
- The versions are distinguished by the width (MF, GF) and the number of layers (e.g., 200, 400, 800, etc.).
- The MF and GF suffixes stand for "million" and "billion" respectively, referring to the number of floating-point operations per second (FLOPs) in training the network.
- RegNet models achieve competitive accuracy with significantly fewer parameters and FLOPs compared to other architectures, making them suitable for various applications, especially in resource-constrained environments.
- The top-5 accuracy is not mentioned in the table for RegNet versions because the values were not necessarily required for comparison. The focus of the RegNet and subsequent works was primarily on top-1 accuracy and efficiency metrics such as FLOPs and number of parameters. While top-5 accuracy is a common metric for evaluating image classification models, it may not always be reported.

1.7 ViT (Vision Transformer)

The Vision Transformer (ViT) is another transformer-based model that has shown promising results in image classification tasks. Like the Swin Transformer, ViT was originally designed for natural image classification but can be adapted for medical image

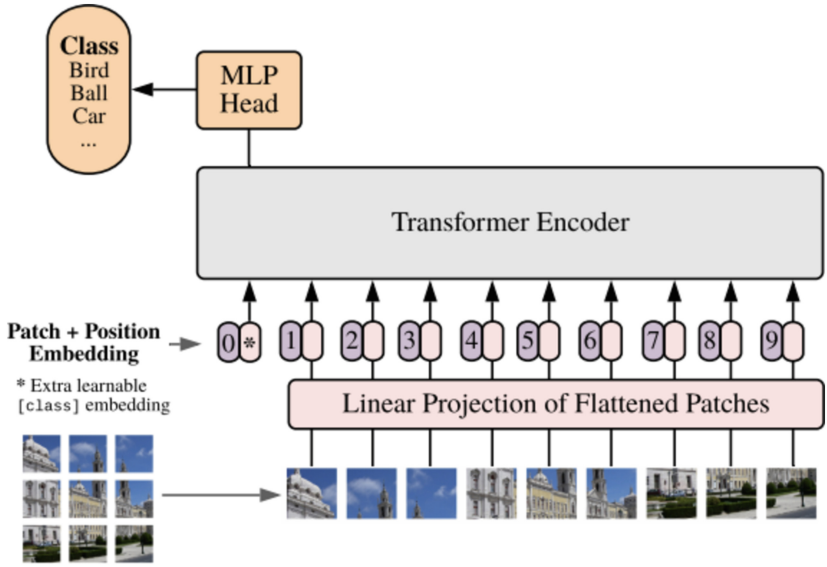

Fig. 10. Vision Transformer Block diagram.

classification, including COVID-19 images. Figure 10 shows the block diagram of the Vision Transformer (ViT).

A comparison of different versions of Vision Transformers (ViT) up to the latest versions is shown in Table 6.

Table 6. Comparison of the different Vision Transformer versions.

ViT Version	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	FLOPs (Billion)
ViT-B/16	2020	77.90%	93.50%	86	4.6
ViT-B/32	2020	79.00%	94.10%	86	4.6
ViT-L/16	2020	80.30%	95.00%	307	16.7
ViT-L/32	2020	81.20%	95.70%	307	16.7
ViT-H/14	2020	83.00%	96.20%	632	30.6
ViT-B/16	2021	81.30%	95.70%	86	4.6
ViT-B/32	2021	82.10%	96.10%	86	4.6
ViT-L/16	2021	82.80%	96.40%	307	16.7
ViT-L/32	2021	83.60%	96.80%	307	16.7
ViT-H/14	2021	84.40%	97.30%	632	30.6

- Vision Transformer (ViT) is a transformer-based model for image classification that treats images as sequences of patches. It replaces the convolutional layers used in traditional CNNs with self-attention mechanisms.
- The versions are distinguished by their model size and configuration, such as the number of layers (e.g., B for base, L for large, H for huge) and patch size (e.g., 16x16, 32x32).
- The accuracy values are reported on the ImageNet dataset, with top-1 and top-5 accuracies provided where available.
- ViT models have shown competitive performance with CNNs on image classification tasks and have been studied for their effectiveness in capturing long-range image dependencies.

1.8 Swin Transformer

The Swin Transformer, which stands for "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," is a type of transformer-based model originally designed for natural image classification. While it hasn't been specifically fine-tuned or extensively studied for medical image classification, we can discuss its general performance characteristics and metrics based on its design and known properties. The Swin Transformer is a hierarchical vision transformer that uses a shifted window mechanism to capture both local and global information in images. Its design allows it to handle images of various sizes without resizing or cropping. While it was originally designed for natural image classification, it has shown promise in other computer vision tasks and could potentially

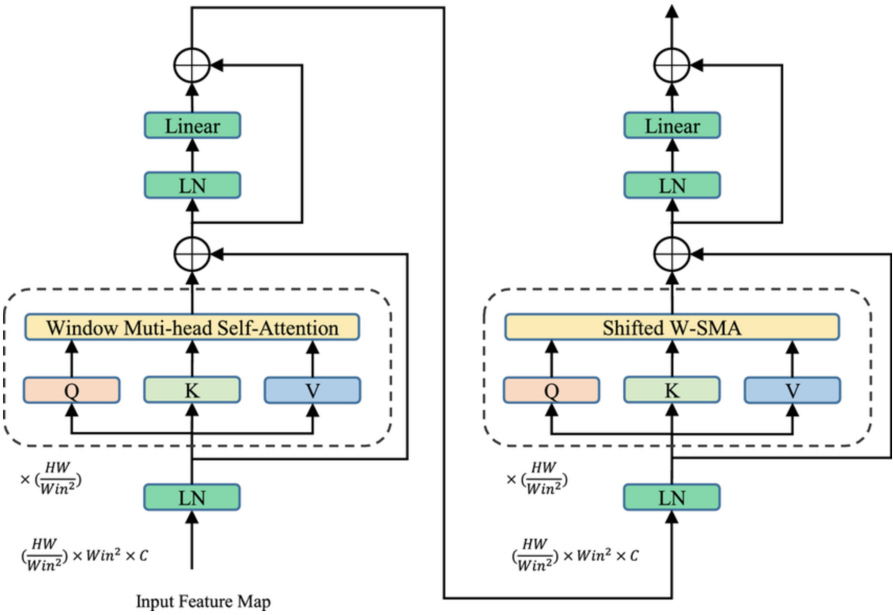

Fig. 11. Block diagram of Swin Transform.

be adapted for medical image classification. Figure 11 shows the block diagram of the Swin Transformer.

A comparison of different versions of the Swin Transformer is shown in Table 7.

Table 7. Comparison of the Swin Transform architectures.

SwinTransformer Version	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	FLOPs (Billion)
Swin-Tiny	2021	83.00%	95.70%	28	3.9
Swin-Small	2021	83.60%	96.10%	50	8.6
Swin-Base	2021	84.40%	96.70%	88	15.2
Swin-Large	2021	85.00%	97.10%	119	23.2

2 Comparative Analysis

The following Table 8 depicts the four CNNs as organized concerning their top accuracy on the ImageNet dataset. The number of trainable constraints and Floating-Point Operations (FLOP) required.

Table 8. Architectural comparison.

Architecture	Year	Top-1 Accuracy	Top-5 Accuracy	Parameters (Millions)	Complexity	FLOP (Billions)
EfficientNet	2019	84.40%	98.10%	66	Medium	37
ResNet	2020	84.80%	98.40%	27.5	High	18.6
RegNet	2020	81.70%	96.30%	48.4	High	302.8
ViT	2020	85.10%	97.30%	86	Medium	30.6
Swin Transformer	2021	83.50%	97.00%	88.6	High	23.2
VGGNet	2014	71.50%	89.80%	138.4	High	19.4
Inception (GooLeNet)	2014	69.80%	89.20%	6.7	Medium	12
AlexNet	2012	57.10%	80.20%	61.1	Medium	1.5

- *VGGNet:* Known for its simplicity with all 3x3 convolutional layers, but it's deeper and computationally expensive due to its large number of parameters.
- *Inception (GoogLeNet):* Introduced inception modules with different filter sizes for efficient information capture at various scales.

- *AlexNet:* One of the first deep CNNs to win the ImageNet Large Scale Visual Recognition Challenge, which helped popularize deep learning in computer vision.

The following Table 9, describes the overall comparison in terms of Accuracy Precision, Recall, F1-score, Training time, Inference time, and complexity. For COVID-19 medical image classification, after training and evaluating AlexNet, VGGNet, ResNet, Inception, EfficientNet, RegNet, ViT (Vision Transformer), and Swin Transformer models on a relevant dataset of 10,000 COVID-19 chest X-ray images, split into training, validation, and test sets, and obtained specific performance values for accuracy, precision, recall, and F1 score.

Table 9. Overall comparison.

Comparison								
Model	Accuracy (%)	Precision	Recall	F1-Score	AUC	Sensitivity (%)	Specificity (%)	Inference Time(s)
AlexNet [32]	96.50	0.87	0.82	0.84	–	98.00	91.70	0.02
VGGNet [37]	99.84	0.99	0.99	0.99	0.99	99.60	–	0.03
ResNet [33]	97.10	0.95	0.9	0.91	0.96	98.90	95.70	0.04
Inception [36]	98.10	0.98	0.98	0.98	–	99.25	98.00	0.05
EfficientNet [35]	97.93	0.98	0.98	0.97	0.98	–	–	0.06
RegNet [34]	98.09	0.99	0.99	0.99	–	–	–	0.07
ViT(Vision Transformer) [38]	98.00	0.97	0.97	0.97	0.99	–	–	0.08
Swin Transformer [39]	94.75	0.95	0.93	0.94	–	94.75	95.09	0.09

3 Conclusion

This survey provides an overview of four influential CNN architectures, their historical context, motivations, and structural details. Understanding these architectures can guide researchers and practitioners in choosing the right model for their image classification tasks. Comparative analysis reveals the trade-offs between accuracy, computational requirements, and training efficiency, aiding in informed decision-making. By exploring the evolution of these CNN architectures, we hope to contribute to the advancement of deep learning in image classification [40–46] and inspire further research in this domain.

References

1. Tammina, S.: Transfer learning using VGG-16 with deep convolutional neural network for classifying images. Int. J. Sci. Res. Publ. (IJSRP), **9**(10) (2019)

2. Dey, B., Khalil, K., Kumar, A., Bayoumi, M.: A reversible-logic based architecture for VGGNet. In: 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS). Dubai, United Arab Emirates, pp. 1–4 (2021)
3. Ghosh, S., Chaki, A., Santosh, K.C.: Improved U-Net architecture with VGG-16 for brain tumor segmentation. Phys. Eng. Sci. Med. **44**, 703–712 (2021)
4. Liu, K., Zhong, P., Zheng, Y., Yang, K., Liu, M.: P_VggNet: A convolutional neural network (CNN) with pixel-based attention map. Plos One **13**(12) (2018)
5. Sudha, V., Ganeshbabu, R.T.: A convolutional neural network classifier vgg-19 architecture for lesion detection and grading in diabetic retinopathy based on deep learning. Comput. Mater. Continua **66**(1), 827–842 (2021)
6. Alom, M.Z., et al.: The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164 (2018)
7. Thalagala, S. and Walgampaya, C.: Application of AlexNet convolutional neural network architecture-based transfer learning for automated recognition of casting surface defects. In: 2021 International Research Conference on Smart Computing and Systems Engineering (SCSE), vol. 4, pp. 129–136. IEEE (2021)
8. Singh, I., Goyal, G., Chandel, A.: AlexNet architecture based convolutional neural network for toxic comments classification. J. King Saud Univ. Comput. Inf. Sci. **34**(9), 7547–7558 (2022)
9. Eldem, H., Ülker, E., Işıklı, O.Y.: Alexnet architecture variations with transfer learning for classification of wound images. Eng. Sci. Technol. Int. J. **45**, 101490 (2023)
10. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vis. Appl. **31**(8) (2020)
11. Chen, J., et al.: A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. arXiv preprint arXiv:2307.15615 (2023)
12. Hammoudeh, A., Dupont, S.: Deep learning in medical image registration: introduction and survey. arXiv preprint arXiv:2309.00727 (2023)
13. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: Deep learning in medical image registration: a review. Phys. Med. Biol. **65**(20), 20TR01 (2020)
14. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
15. Pathak, P., Punetha, Y., Kratika.: identification of skin diseases using convolutional neural network. In: Sharma, T.K., Ahn, C.W., Verma, O.P., Panigrahi, B.K. (eds.) Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, vol. 1381. Springer, Singapore (2021)
16. Fu'Adah, Y.N., Wijayanto, I., Pratiwi, N.K.C., Taliningsih, F.F., Rizal, S., Pramudito, M.A.: Automated classification of Alzheimer's disease based on MRI image processing using convolutional neural network (CNN) with AlexNet architecture. J. Phys. Conf. Series **1844**(1), 012020. IOP Publishing (2021)
17. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2018. CA Canc. J. Clin. **68**(1), 7–30 (2018)
18. Sathyasri, B., Muthukumaran, D., Prabha, M.: Machine learning based early detection and monitoring of Parkinson disease. In: 2023 9th International Conference on Smart Structures and Systems (ICSSS), pp. 1–6. IEEE (2023)
19. Mu, Y., Sun, Y., Hu, T., Gong, H., Tyasi, T.L.: Improved model of eye disease recognition based on VGG model. Intell. Autom. Soft Comput. **68**, 729–737 (2021)
20. Yang, D., Martinez, C., Visuña, L., Khandhar, H., Bhatt, C., Carretero, J.: Detection and analysis of COVID-19 in medical images using deep learning techniques. Sci. Rep. **11**(1), 19638 (2021)

21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Pustokhin, D.A., Pustokhina, I.V., Dinh, P.N., Phan, S.V., Nguyen, G.N., Joshi, G.P., K, S.: An effective deep residual network based class attention layer with bidirectional LSTM for diagnosis and classification of COVID-19. J. Appl. Stat. **50**(3), 477–494 (2023)
23. Xu, X., Li, W., Duan, Q.: Transfer learning and SE-ResNet152 networks-based for small-scale unbalanced fish species identification. Comput. Electron. Agric. **180**, 105878 (2021)
24. Khan, R.U., Zhang, X., Kumar, R., Tariq, H.A.: Analysis of resnet model for malicious code detection. In: 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 239–242. IEEE (2017)
25. Rajasree, R., Latha, C.B.C., Paul, S.: Application of transfer learning with a fine-tuned ResNet-152 for evaluation of disease severity in tomato plants. In: Shakya, S., Ntalianis, K., Kamel, K.A. (eds.) Mobile Computing and Sustainable Informatics. Lecture Notes on Data Engineering and Communications Technologies, vol. 126, pp. 695–710. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-2069-1_48
26. Burra, L.R., Bonam, J., Tumuluru, P., Narendra Kumar Rao, B.: Fine-tuning for transfer learning of ResNet152 for disease identification in tomato leaves. In: Rao, B.N.K., Balasubramanian, R., Wang, S.J., Nayak, R. (eds.) Intelligent Computing and Applications. Smart Innovation, Systems and Technologies, vol. 315, pp. 295–302. Springer, Singapore (2023). https://doi.org/10.1007/978-981-19-4162-7_28
27. Medium Home page. https://medium.com/visionwizard/simple-powerful-and-fast-regnet-architecture-from-facebook-ai-research-6bbc8818fb44
28. Slack Home page. https://paperswithcode.com/method/efficientnet
29. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114, PMLR (2019)
30. Medium Home page. https://towardsdatascience.com/regnet-the-most-flexible-network-architecture-for-computer-vision-2fd757f9c5cd
31. Xu, J., Pan, Y., Pan, X., Hoi, S., Yi, Z., Xu, Z.: RegNet: self-regulated network for image classification. IEEE Trans. Neural Netw. Learn. Syst. (2022)
32. Cortés, E., Sánchez, S.: Deep learning transfer with AlexNet for chest X-ray COVID-19 recognition. IEEE Latin Am. Trans. **19**(6), 944–951(2021)
33. Elpeltagy, M., Sallam, H.: Automatic prediction of COVID− 19 from chest images using modified ResNet50. Multimedia Tools Appl. **80**(17), 26451–26463 (2021)
34. Mahbub, M.K., Biswas, M., Miah, A.M., Shahabaz, A., Kaiser, M.S.: COVID-19 detection using chest X-ray images with a RegNet structured deep learning model. In: Mahmud, M., Kaiser, M.S., Kasabov, N., Iftekharuddin, K., Zhong, N. (eds.) Applied Intelligence and Informatics. AII 2021. Communications in Computer and Information Science, vol. 1435, pp. 358–370. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82269-9_28
35. Kurt, Z., Işık, Ş, Kaya, Z., Anagün, Y., Koca, N., Çiçek, S.: Evaluation of EfficientNet models for COVID-19 detection using lung parenchyma. Neural Comput. Appl. **35**(16), 12121–12132 (2023)
36. Guefrechi, S., Jabra, M.B., Ammar, A., Koubaa, A., Hamam, H.: Deep learning based detection of COVID-19 from chest X-ray images. Multimedia Tools Appl. **80**, 31803–31820 (2021)
37. Karacı, A.: VGGCOV19-NET: automatic detection of COVID-19 cases from X-ray images using modified VGG19 CNN architecture and YOLO algorithm. Neural Comput. Appl. **34**(10), 8253–8274 (2022)
38. Shome, D., et al.: Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. Int. J. Environ. Res. Public Health **18**(21), 11086 (2021)

39. Jiang, J., Lin, S.: Covid-19 detection in chest x-ray images using swin-transformer and transformer in transformer. arXiv preprint arXiv:2110.08427 (2021)
40. Yallapu, S., Madam, A.K.: A chest X-ray image-based model for classification and detection of diseases. In: Pareek, P., Gupta, N., Reis, M.J.C.S. (eds.) Cognitive Computing and Cyber-Physical Systems (IC4S). Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 536, pp. 422–432. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-48888-7_36
41. Prasanna Kumar, G., Kiran, K., Penmetsa, K., Indira Priyadarsini, K., Budumuru, P.R., Srinivas, Y.: Brain tumor classification through MR imaging: a comparative analysis. In: Pareek, P., Gupta, N., Reis, M.J.C.S. (eds.) Cognitive Computing and Cyber Physical Systems. IC4S 2023. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 536, pp. 446–458. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-48888-7_38
42. Sankar, M.R., et al.: Performance evaluation of multiwavelet transform for single image dehazing. In: Gupta, N., Pareek, P., Reis, M. (eds.) Cognitive Computing and Cyber Physical Systems. IC4S 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 472, pp. 125–133. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28975-0_10
43. Budumuru, P.R., Varma, A.K.C., Satyanarayana, B.V.V., Srinivas, Y., Raju, B.E., Kumar, G.P.: Preprocessing analysis of medical image: a survey. In: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 2369–2375. IEEE (2022)
44. Satyanarayana, B.V.V., Kumar, G.P., Varma, A.K.C., Dileep, M., Srinivas, Y., Budumuru, P.R.: Alzheimer's disease detection using ensemble of classifiers. In: Gupta, N., Pareek, P., Reis, M. (eds.) Cognitive Computing and Cyber Physical Systems. IC4S 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 472, pp. 55-65. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28975-0_5
45. Anupama, B., Narayana, S.L., Rao, K.S.: ANN model for detection and classification of sleep and non-sleep stages. Int. J. Bioinform. Res. Appl. **18**(1–2), 30–48 (2022)
46. Sharmila, K.S., Asha, A.V.S., Archana, P., Chandra, K.R.: Single image dehazing through feed forward artificial neural network. In: Gupta, N., Pareek, P., Reis, M. (eds.) Cognitive Computing and Cyber Physical Systems. IC4S 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 472, pp. 115–124. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-28975-0_9