International Journal of
Image Processing and
Pattern Recognition

ISSN: 2456-6985
Volume 12, Issue 1, 2026
January—June
DOI (Journal): 10.37628/IJIPPR

https://journalspub.com/journal/ijippr/

Review IJIPPR

# A Comprehensive Review of Convolutional Neural Network Architectures and Evolution

Swarnali Kundu[1], Biswasri Datta[1], Tousif Parvej[1], Pritam Pal[1], Fakruddin Ali Ahmed[2,*]

***Abstract***

*Convolutional Neural Networks (CNNs) have become a foundational deep learning framework in computer vision because they can automatically extract layered, increasingly complex features directly from raw image data. CNNs use convolutional, pooling, and activation layers to extract spatial patterns from basic edges to intricate object components, drawing inspiration from the human visual cortex. An overview of CNNs' basic architecture is given in this study, along with an explanation of important elements such kernels, convolution processes, pooling strategies, activation functions, and fully connected layers. To show how deep learning architectures have evolved in terms of depth, computational efficiency, and feature extraction capabilities, classic CNN models such as LeNet, AlexNet, VGGNet, and GoogLeNet are examined. Advances in large-scale picture categorization and recognition problems have been greatly aided by these structures. CNNs still have drawbacks despite their effectiveness, including high processing costs, the need for large labeled datasets, and interpretability issues. In order to serve real-time and embedded applications, future developments are anticipated to concentrate on lightweight architectures, increased model transparency, and improved training efficiency. All things considered, CNNs continue to be crucial to contemporary artificial intelligence research because they let machines to process and comprehend visual data with ever-increasing precision and resilience.*

**Keywords:** CNN architecture, CNN models, convolutional neural network (CNN), deep learning, neural network

## INTRODUCTION

Convolutional Neural Networks (CNNs) have become breakthrough technology, driving progress in computer vision and image processing. CNNs demonstrated unparalleled performance in object detection, facial recognition, and classification by leveraging layered architectures that mimic the visual experience of the human brain. Through their accomplishments, academic research has advanced and innovation has been sparked in several industries: healthcare, automotive, and security. This paper explores CNNs' basic concepts, development, topologies, and wide range of applications for providing readers a comprehensive idea of how CNNs function and why they remain crucial to deep learning methodologies.

***Author for Correspondence**
Fakruddin Ali Ahmed
E-mail: fakruddinahmed85@gmail.com

[1]Student, Department of Information Technology, B. P. Poddar Institute of Management & Technology, Kolkata, West Bengal, India
[2]Assistant Professor, Department of Information Technology, B. P. Poddar Institute of Management & Technology, Kolkata, West Bengal, India

Received Date: October 14, 2025
Accepted Date: December 22, 2025
Published Date: February 18, 2026

**Citation:** Swarnali Kundu, Biswasri Datta, Tousif Parvej, Pritam Pal, Fakruddin Ali Ahmed. A Comprehensive Review of Convolutional Neural Network Architectures and Evolution. International Journal of Image Processing and Pattern Recognition. 2026; 12(1): 32–40p.

## LITERATURE REVIEW

Usually, a convolutional neural network is made up of several layers that are arranged according to their functions. In particular, CNN is inspired by the cortical area's anatomy. The cortical region of a

 

mammal was supported by the structural model. There are tiny cell clusters in the brain that are responsive to certain regions of the visual field. In 1962, Hubel, and Wiesel conducted an experiment that expanded on this idea [1]. Fukushima designed the Neocognitron in 1980, which is the main network that uses the graded organization of neurons and the type of local connection to change the image. According to Fukushima, options for the translational data invariance are produced after some parameters of a nerve cell are altered to a space within a completely new place of the preorder layer [2]. ConvNets are considered to have their roots in the neocognitron of 1980. Then, LeCun et al. (1989) and LeCun et al. (2002) [3–4] designed and created the CNN framework by creating seven learned layers, which included four convolutional and pooling layers, and then LeNet-5, a three-layer artificial neural network with three fully connected layers. LeNet-5 was trained using the algorithmic backpropagation program [5], allowing it to recognize patterns from raw pixels, eliminating additional feature extraction techniques. LeNet-5 was used to categorize printed numbers. However, this approach did not perform well for difficult problems due to insufficient computer power and training data.

A CNN model developed by Krizhevsky et al. (2017) [6] was successful in lowering the mistake rate on ILSVRC competition by a large margin [7]. A deep CNN design known as AlexNet [6] was proposed by Krizhevsky et al. and is a crucial improvement in image classification tasks. With eight learned layers this architecture is regarded as a substantial variation of LeNet. Although AlexNet featured a broader structure than LeNet-5, it initially performed on par with the classic LeNet-5. However, over time, AlexNet outperformed the earlier ConvNets model in computer vision by employing supervised learning while maintaining the model simplicity [3, 8–10]. We have examined the convolutional neural network's basic architecture in this review.

## CNN ARCHITECTURE

A deep Convolutional Neural Network (CNN) consists of several interconnected layers that gradually learn and extract features at different levels of abstraction from input data, such images. While deeper layers capture more intricate, high-level information like forms and object pieces, the early layers are in charge of recognizing low-level elements like edges and textures. The overall architecture of a CNN is outlined in Figure 1, with the succeeding sections offering a detailed examination of the functional contributions of each layer [10].

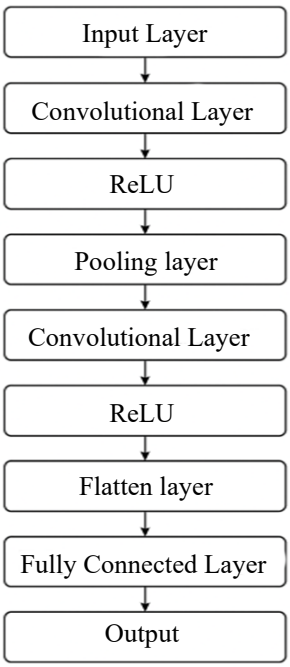


**Figure 1.** Conceptual model of CNN [4].

 

Several different and distinct kinds of layers make up a standard Convolutional Neural Network (CNN), each of which is intended to carry out a distinct task during the feature learning and representation process [11].

**Convolutional Layer**

The primary function of convolutional layer in CNN is to extract features from the input data. It functions by applying a number of learnable filters, called kernels, that traverse the input and perform various convolution operations. These operations allow the network to detect spatial patterns, such as the edges, textures, and other structural features [12].

**Kernel**

A kernel is a small, trainable numerical matrix composed of numerical values, each representing a trainable weight. At the start of training, the network's weights are generally assigned random values, although several specialized initialization techniques may also be used.

During training, backpropagation iteratively updates the weights, allowing the kernel to progressively learn meaningful feature representations from the input data. Figure 2 illustrates an example of a two-dimensional convolutional filter [13].

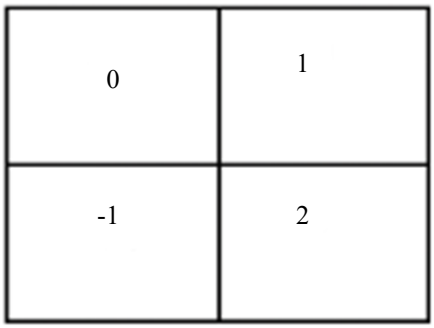


**Figure 2.** Kernel of size 2 × 2.

**Convolution Operation**

To understand the convolution operation in CNNs, it is important to consider the input data format. Unlike traditional neural networks, which accept inputs as one-dimensional vectors, CNNs are modeled in image data handling for its multidimensional form. For example, an RGB image has three channels: Red, Green, and Blue as shown in Figure 3. In contrast, a grayscale image only has one channel and can be represented as a two-dimensional matrix as shown in Figure 4. The convolution operation performed in this layer is shown in Figure 5.

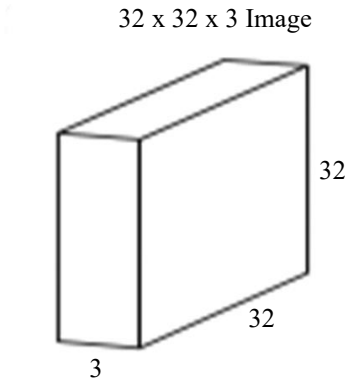


**Figure 3.** RGB image.

 

| | | | |
|---|---|---|---|
| 1 | 0 | -2 | 1 |
| -1 | 0 | 1 | 2 |
| 0 | 2 | 1 | 0 |
| 1 | 0 | 0 | 1 |

**Figure 4.** Gray-Scale image of size 4 × 4.

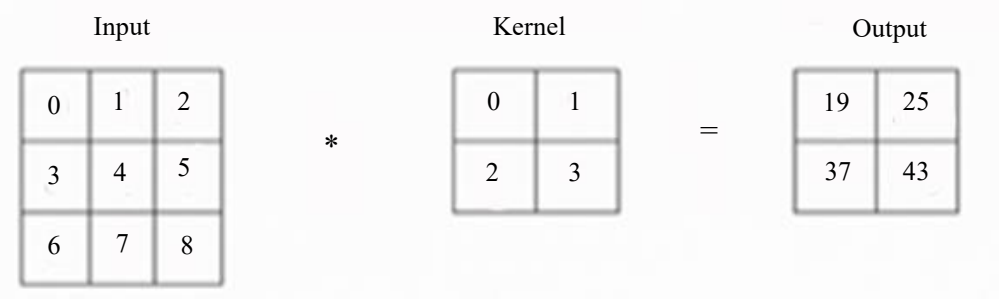


**Figure 5.** Illustrating of convolution operation.

**Pooling Layer**

Convolutional Neural Networks (CNNs) need a pooling layer for minimizing the feature maps' spatial dimensions while keeping the most crucial data intact. In order to aggregate or summarize data within small, localized sections of the input feature map, it usually employs techniques like max pooling or average pooling. This down sampling process reduces the amount of parameters and computations in the network, increases the model's resistance to minute translations or distortions in the input data, and helps prevent overfitting.

***Pooling Methods***

Pooling is an essential procedure of "Convolutional Neural Networks (CNNs)" that reduces the feature map's spatial dimensions, preserving important characteristics while reducing the computational cost and parameter count. It also renders the network invariant to small translations, distortions, and scaling of inputs.

Three commonly used pooling methods:

- Max Pooling.
- Mean Pooling.
- Average Pooling.

In CNN architectures, max pooling is the most popular of them. By choosing the highest value from each local feature map region, it successfully keeps the most noticeable feature in that area. Figure 6 shows an example of this procedure. Pooling improves the model's resilience to changes in shape, size, and orientation while simultaneously making the feature map simpler. Figure 6 illustrates max pooling, where the highest value in each region is selected.

 

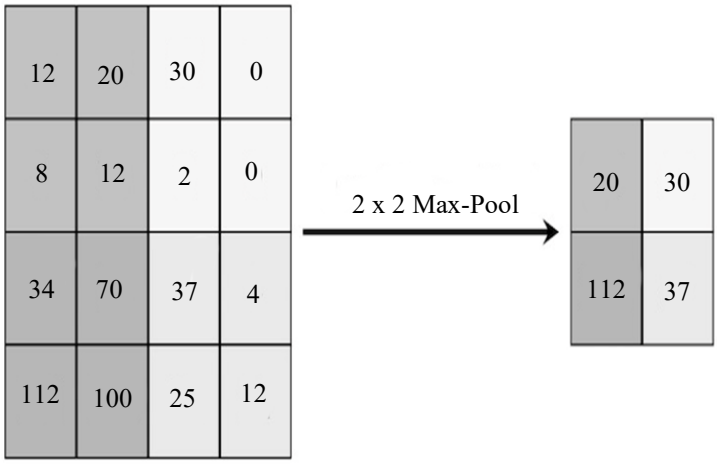


**Figure 6.** Illustrating of max pooling operation.

**Activation Functions**

An activation function is a mathematical mechanism that determines a neuron's level of activation. By introducing non-linearity into the network, it enables the model to identify complex patterns and relationships within the input data [2]. Without activation functions, a neural network – regardless of its depth – would reduce to a simple linear model.

Generally used activation functions are:

- Rectified Linear Unit(ReLU).
- Leaky ReLU.
- Sigmoid.

ReLU's simplicity and computing efficiency make it the most commonly utilized activation function in CNNs. The following equation defines how it changes input data: It sets all negative values to zero while leaving positive ones unaltered. ReLU's fundamental benefit is that it keeps deep networks performing well while speeding up convergence during training.

**Fully Connected Layers (FC Layers)**

Fully Connected (FC) layers, also known as dense layers, are a type of feedforward artificial neural network component that follows the traditional structure of a Multi-Layer Perceptron (MLP). In CNNs, FC layers are typically placed after the convolutional and pooling layers. They take the high-level feature maps as input – flattened into a one-dimensional vector – and process them to produce the final result of the model, such as class scores in a classification task. As shown in Figure 7, the feature maps from the preceding layers are first flattened. The flattened layer is passed through fully connected layers, which learn complex combinations of features to generate the final predictions.

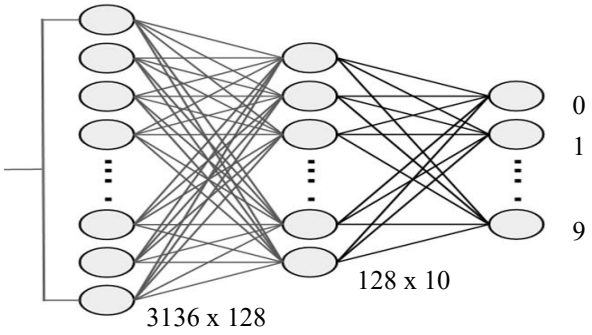


**Figure 7.** The architecture of fully connected layers.

 

## CNN MODELS

Classic CNN models such as "LeNet", "AlexNet", "VGG–16", and "VGG–19" – have demonstrated strong performance across various image classification tasks and continue to serve as foundational models in deep learning research. These architectures are great sources of inspiration for creating new CNN models since they are generally accepted to be dependable and efficient. In these architectures, recurring convolutional and pooling layer combinations are a frequently used structural pattern. Typical setups consist of:

- Convolution – Convolution – Pooling – Convolution – Convolution – Pooling
- Convolution – Pooling – Convolution – Pooling

These patterns allow for progressively deeper feature extraction while reducing spatial dimensions, ensuring a balance between computational efficiency and learning capability.

### LeNet

Originally created for the purpose of recognizing handwritten characters, LeNet is among the first and most significant Convolutional Neural Network (CNN) architectures. The architecture has convolutional layers, pooling layers, fully connected layers and finally softmax classifier for output prediction.

### VGGNet

VGGNet is a deep convolutional neural network that begins with an input image of 224 × 224 × 3 and processes it through many stacked layers of 3 × 3 convolutions, each activated by ReLU. These convolutional stages gradually learn visual patterns, starting from basic edges and textures and progressing to more abstract structures. After every few convolution layers, a 2 × 2 max pooling operation reduces the height and width of the feature maps while keeping the most important information. Through this sequence, the network produces increasingly compact feature representations, eventually reaching a 7 × 7 × 512 feature map. This final feature map is then flattened and passed into two large fully connected layers with 4096 units, which function like a traditional classifier. The last fully connected layer outputs 1000 scores, one for each possible object category. VGGNet's strength lies in its straightforward yet deep design – using only small filters, uniform layer structures, and consistent pooling – allowing it to capture rich visual details and achieve strong classification performance. Its architecture is displayed in Figure 8.

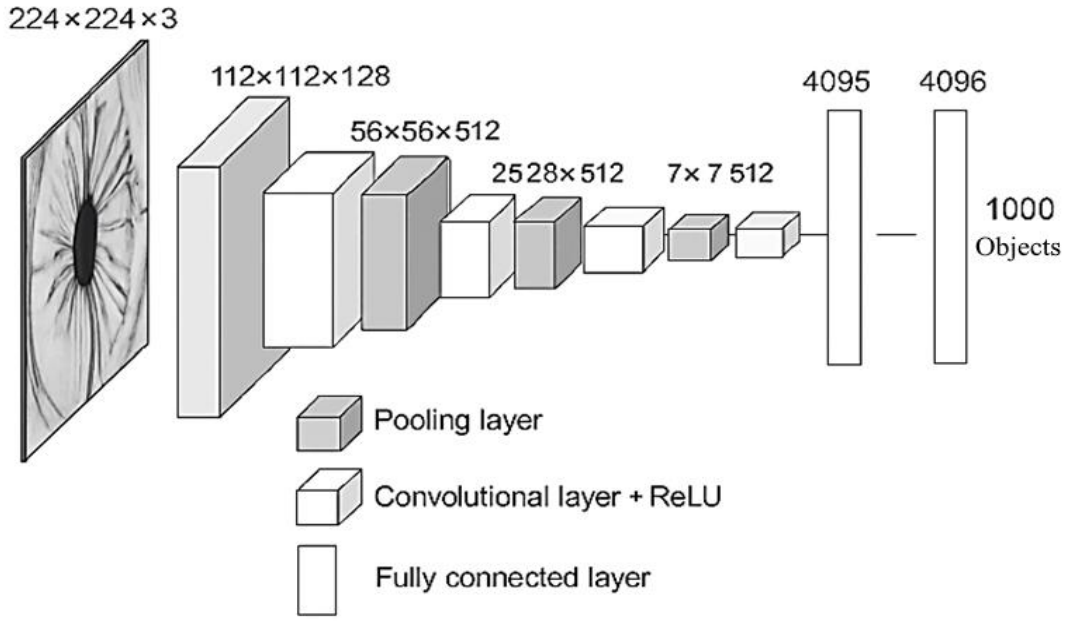


**Figure 8.** VGGNet.

 

**AlexNet**

The ground-breaking CNN architecture AlexNet made substantial strides in deep learning and won the ILSVRC 2012. It consists of five convolutional layers, five max pooling layers in between, and then three fully connected layers.The network continuously employs the ReLU activation function, which adds non-linearity without raising computing complexity, speeding up training. In order to capture information at various spatial scales, AlexNet uses a range of filter sizes, including 3 × 3, 5 × 5, and 11 × 11, in contrast to later models that mostly rely on small filters. Its deep structure and multi-scale methodology helped it achieve remarkable results and leave a lasting impression on CNN design. Its structural layout is shown in Figure 9.

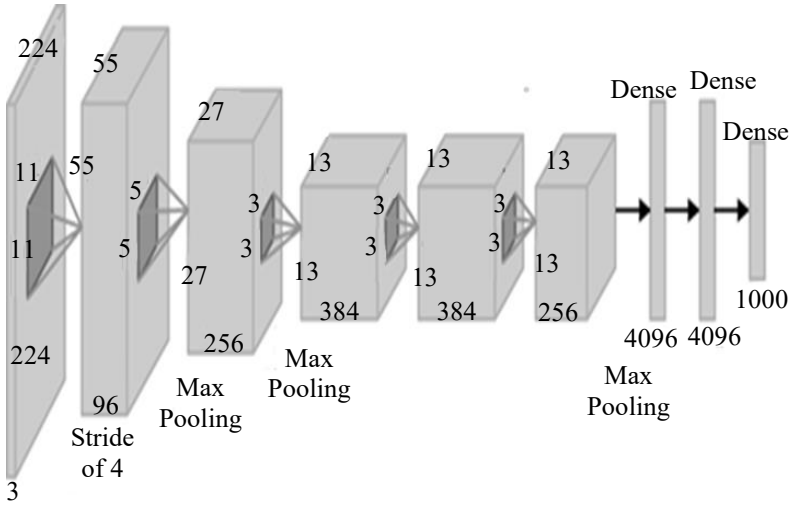


**Figure 9.** AlexNet.

**GoogLeNet**

Google researchers created GoogLeNet, which, thanks to its innovative and effective architecture, won first place in the ILSVRC competition. The Inception module, at the heart of its design, applies several convolutional layers in parallel, each with a distinct filter size (e.g., 1 × 1, 3 × 3, and 5 × 5). To create a single feature map, the outputs of these parallel layers are concatenated after a pooling procedure. This approach improves the network's capacity to identify intricate patterns while maintaining comparatively low computational costs by allowing it to learn features at different resolutions and scales inside the same layer.

Auxiliary classifiers are added to GoogLeNet's intermediary layers to increase training effectiveness and decrease overfitting. These extra branches help the network learn more robust and discriminative features by acting as regularizers. Figure 10(a) shows a simple Inception module, while Figure 10(b) presents a dimension-reduced version using 1 × 1 bottleneck layers.

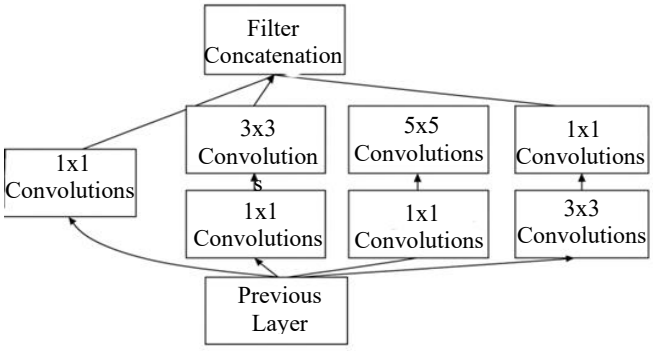


**Figure 10(a).** Simple inception module.

 

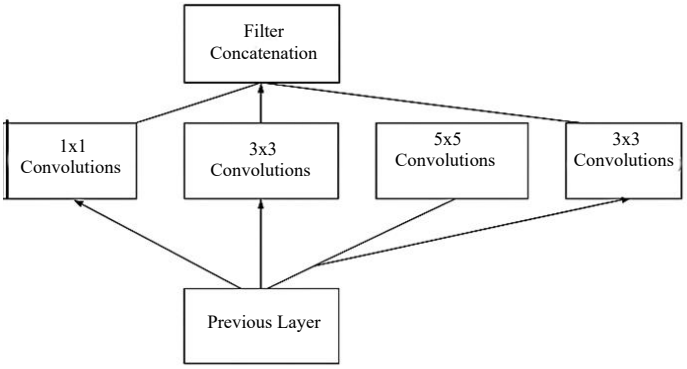


**Figure 10(b).** Reduced dimension reduction inception module.

Although certain traditional CNN architectures are briefly introduced in this part, you will gain a deeper grasp of deep learning in computer vision by investigating more sophisticated models, such as MobileNet, ResNet, R–CNN and Fast R–CNN.

## CONCLUSION

From basic digit identification models to complex architectures that can manage extensive, real-world image classification tasks, Convolutional Neural Networks (CNNs) have advanced. CNNs have continuously raised the bar for computer vision performance through improvements in depth, parameter efficiency, and feature extraction methods. Enhancing model interpretability, utilizing transfer learning, and creating scalable, lightweight architectures appropriate for real-time processing and embedded system deployment should be the top priorities of future research.

## REFERENCES

1. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962;160:106–154.
2. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 1980 Apr;36(4):193–202.
3. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989 Dec;1(4):541–551.
4. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 2002 Aug;86(11):2278–2324.
5. Karayiannis N, Venetsanopoulos AN. Artificial neural networks: Learning algorithms, performance evaluation, and applications. 1st ed. New York (NY): Springer Science & Business Media; 2013. p. 1–450.
6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017 May;60(6):84–90.
7. Li LJ, Su H, Lim Y, Fei-Fei L. Object bank: An object-level image representation for high-level visual recognition. Int J Comput Vis. 2014 Mar;107(1):20–39.
8. Jarrett K, Kavukcuoglu K, Ranzato MA, LeCun Y. What is the best multi-stage architecture for object recognition? In: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision; 2009 Sep 29–Oct 2; Kyoto, Japan. IEEE; 2009. p. 2146–2153.
9. Krizhevsky A, Hinton G. Convolutional deep belief networks on CIFAR-10. Unpublished manuscript. 2010 Aug;40(7):1–9.
10. Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. In: Proceedings of the 2011 International Conference on Computer Vision; 2011 Nov 6–13; Barcelona, Spain. IEEE; 2011. p. 2018–2025.
11. Shin JS, Ma J, Choi SJ, Kim S, Hong M. Development of a deep learning model for predicting speech audiometry using pure-tone audiometry data. Appl Sci (Basel). 2024 Oct 15;14(20):9379.
12. Shawky NE. Convolutional neural network and its applications in artificial intelligence. J ACS Adv Comput Sci. 2021 Jun;12(1):10–26.

 

13. Balas VE, Kumar R, Srivastava R, editors. Recent trends and advances in artificial intelligence and internet of things. 1st ed. Cham (CH): Springer International Publishing; 2020. p. 1–520.

 