Image recognition using deep learning: a review

Osama M. Hassan, Ashraf A. Gouda, Mohammed Abdel Razek
Department of Mathematics and Computer Science, Faculty of Science, Al-Azhar University, Nasr City, Egypt

Article Info

Article history:

Received Sep 5, 2024
Revised Aug 9, 2025
Accepted Oct 14, 2025

Keywords:

AI
CNN
Deep learning
Image recognition
Machine learning

ABSTRACT

This paper presents a comprehensive review of recent advancements in image recognition, with a focus on deep learning (DL) techniques. Convolutional neural networks (CNNs), in particular, have significantly transformed this domain, enabling substantial improvements in both accuracy and efficiency across diverse applications. The review explores state-of-the-art methods, highlighting their practical implementations and the progress achieved. It also addresses key challenges such as data scarcity and model interpretability, offering perspectives on emerging opportunities and future directions. By synthesizing current trends with forward-looking insights, the paper aims to serve as a valuable resource for researchers and practitioners seeking to navigate and contribute to the evolving landscape of image recognition. Moreover, the paper examines critical challenges that persist in the field, such as transfer learning, data augmentation, and explainable artificial intelligence (AI) approaches. By synthesizing current trends with emerging innovations, the review not only maps the trajectory of progress but also highlights future directions and research opportunities. This synthesis aims to provide researchers, developers, and industry practitioners with a solid understanding of the dynamic and rapidly evolving environment surrounding image recognition technologies.

Corresponding Author:

Osama M. Hassan
Department of Mathematics and Computer Science, Faculty of Science, Al-Azhar University
Nasr City 11884, Cairo, Egypt
Email: osama.aldhefeery@gmail.com.

1. INTRODUCTION

Image recognition, sometimes referred to as computer vision or image classification, is an essential component of artificial intelligence (AI), as it involves the automatic recognition and classifying of objects, scenes, or patterns within digital images. It is integral to a variety of applications, such as autonomous vehicles, robotics, medical imaging, surveillance, e-commerce, and multimedia processing [1]. For a long time, accurately and efficiently recognizing and interpreting images has posed a significant challenge in computer science. Traditional image recognition methods depended on manually crafted features and rule-based algorithms, which demanded a great deal of manual labour to extract meaningful information from images. The emergence of machine learning (ML) and deep learning (DL) techniques has revolutionized image recognition, enabling computers to learn directly from raw image data and eliminating the need for explicit feature engineering [2]. This data-driven approach has transformed image recognition by allowing computers to learn representations directly from the raw pixel values of images [3]. The availability of vast datasets, improvements in algorithm development, and increases in processing capacity have all contributed to the considerable evolution of AI over time. The development of intelligent systems has been significantly aided by ML, which has allowed computers to learn and enhance their performance on particular tasks through experience. ML algorithms, particularly supervised learning methods, have been widely used for image recognition tasks. These algorithms learn from a labelled dataset, where images are associated with predefined class labels. Through the training process, they learn to identify patterns and correlations between pixel values and corresponding labels, allowing them to make accurate predictions on unseen images. K-nearest neighbours (KNN), random forests, and support vector machines (SVM) are widely used ML algorithms used for image identification. but their effectiveness is often limited by the need for handcrafted features and complex feature engineering [4], [5]. Within the domain of ML, DL has arisen as a potent subset that has sparked a significant scientific revival in image processing tasks, allowing computers to directly learn intricate patterns and representations from raw image data [6]. DL has dramatically transformed the field of image recognition by enabling remarkable advances in computer vision tasks. Image recognition involves the automatic identification and categorization of objects, scenes, or patterns within digital images. DL techniques, particularly deep neural networks (DNNs), have demonstrated exceptional learning capabilities and extracting intricate features directly from raw image data, resulting in significant advancements in image recognition accuracy and performance [7]. DL has become a potent method for image recognition. DNNs modeled after the structure and function of the human brain, have shown outstanding abilities in learning and extracting complex features directly from raw image data. Models like convolutional neural networks (CNNs) in DL utilize interconnected layers of nodes to gradually learn hierarchical representations of images [8]-[10]. This enables DNNs to capture and comprehend intricate patterns, textures, and spatial relationships within images, resulting in substantial enhancements in image recognition accuracy [11]. The quality of images is essential in DL image recognition, as it significantly impacts the model’s ability to generalize. Before training the model, image preprocessing is conducted to remove irrelevant information, improve the visibility of useful data, and simplify the data. The purpose of this step is to enhance the model’s feature extraction and recognition reliability [12]. Several factors have contributed to the success of DL in image recognition. One key factor is the presence of extensive labelled datasets, like ImageNet, which have allowed DNNs to be trained on a wide variety of images. These datasets provide ample training data for models to learn from and perform well on new images [13]. Furthermore, improvements in computational resources, especially GPUs, have sped up the training and inference processes of DL models, making large-scale image recognition possible [14], [15]. The impact of image recognition extends far beyond academic research and has transformed industries and applications. In autonomous vehicles, image recognition plays a crucial role in object detection, lane detection, and traffic sign recognition. In healthcare, it aids in medical image analysis, disease diagnosis, and treatment planning [16].

In security and surveillance, it assists in face recognition, object tracking, and anomaly detection. Furthermore, image recognition has enabled new possibilities in e-commerce, such as visual search and recommendation systems, and has facilitated innovative multimedia processing techniques [17]. The process of recognizing images using DL involves several main steps. Firstly, a large dataset of desired images is collected, which is obtained from various sources. Then, the collected dataset is used for training the DL model, usually employing CNNs or recurrent neural networks-(RNNs) [18]. The model is taught to identify significant features and patterns in the images, allowing it to categorize and identify various objects or specific important details. During the training process, the model’s parameters are continually adjusted to reduce the disparity between expected and actual labels, improving its capacity to precisely recognize and classify images. The DL model can be used to identify and examine brand-new, invisible images after it has been trained [19].

There are still some problems that have not been solved. Current models are ineffective on lower-quality images, and interpretability is a significant barrier to their adoption. Furthermore, these models need to be developed into frameworks that would allow for a smooth amalgamation of DL with image processing techniques for improving overall performance. This paper intends to fill such gaps, by providing a thorough and detailed overview of the state-of-the-art image recognition methodologies developed through DL. We will present different methods, their advantages and disadvantages, and possible directions for future research for building stronger and more efficient systems with host performances. It is envisioned to synthesize the existing literature into the improved areas and, thus, contribute to the evolving discussion in the area as well as pave the way for future revolutions in image recognition technology.

2. IMAGE RECOGNITION

The process of image processing encompasses a series of actions designed to manipulate and improve digital images in order to extract important information or enhance their visual quality. The basic concept of image recognition is to identify specific features within the image. This recognition process involves three primary stages: image processing, extraction of image features, and image classification [5]. (as depicted in Figure 1).

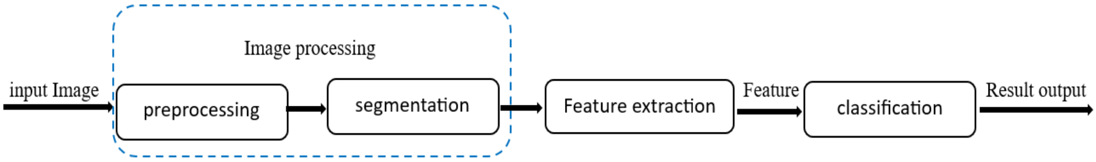

Figure 1. Process of image recognition

The process of utilizing computers to modify images to meet specific criteria for subsequent recognition is referred to as image processing. This procedure is mainly bifurcated into two stages: image preprocessing and image segmentation. Image preprocessing encompasses activities like image restoration and transformation, primarily aimed at eliminating disturbances and noise, enhancing valuable information, and refining object detectability. Moreover, real-time image processing involves re-encoding and compressing the image to decrease algorithmic complexity and enhance computational efficiency. Conversely, image segmentation entails partitioning the identified image into multiple sub-regions, each exhibiting unique attributes and sharing certain similarities in their internal characteristics. The current methods for image segmentation primarily include threshold-based segmentation, edge-based segmentation, and region-based segmentation [20], [21].

3. DEEP LEARNING

Image processing has undergone a revolution thanks to a ML subfield. DL models that is, DNNs are designed to mimic the composition and functionality of the human brain. They are made up of several linked layers of nodes (neurons) that process and modify input data, gradually gaining knowledge of and extracting data’s hierarchical representations [22]. By automatically learning and identifying pertinent features from unprocessed image data, DNNs eliminate the requirement for explicit feature engineering. Image processing tasks such as image recognition, object detection, image segmentation, and image generation have been significantly influenced by DL [23]. CNNs, a popular DL structure, have shown remarkable performance in image recognition by extracting local spatial patterns and features from images. CNNs have been used in many different domains, such as medical imaging, autonomous cars, facial recognition, and satellite imagery analysis [24]-[26]. Several factors have facilitated the development of DL models for image processing. The existence of large-scale labelled datasets, like ImageNet, has permitted researchers to train DNNs on diverse image examples, enabling the models to learn rich representations and generalize effectively to new data. Additionally, progress in computational resources, especially graphical processing units (GPUs), has expedited the training and inference processes of DL models, making large-scale image processing viable. The integration of AI, ML, and DL has transformed the image-processing field [7].

3.1. CNNs

CNNs are a type of DL model specifically designed for image recognition and computer vision tasks. They have proven to be highly effective in various applications, such as image classification, object detection, and image segmentation. CNNs are inspired by the visual processing in the human brain and excel in capturing spatial patterns and hierarchical features in images [27], [28]. A CNN model typically comprises convolutional layers, pooling layers, and fully connected layers. Convolutional layers utilize a set of filters or kernels to extract local spatial patterns and features from the input data. Pooling layers downsample the spatial dimensions to reduce computational complexity and provide translation invariance. Fully connected layers link every neuron in the preceding layer to the following layer, allowing the model to acquire more comprehensive and abstract representations [29], [30].

3.1.1. Structure of CNN:

Typically, a CNN has multiple layers [31]:

- Convolutional layers: these layers perform feature extraction by applying multiple filters (also called kernels) to the input image. Each filter within the convolutional layer identifies distinct patterns in the image, like corners, edges, or textures. The resulting convolutions generate feature maps that depict various attributes of the image.
- Activation layers: following the convolution process, an activation function (typically ReLU – Rectified-Linear Unit) is applied to each element of the feature maps. This gives non-linearity to the model and aids in learning complex patterns.
- Pooling layers: the purpose of the pooling layers is to reduce the spatial dimensions of the feature maps by down-sampling them. A popular approach called “max pooling” selects the largest value inside a

limited region, generally 2x2. This operation reduces computational complexity while preserving crucial information.

- Fully connected layers: these layers use the characteristics that were retrieved from earlier levels to perform categorization. Similar to conventional neural networks, the feature maps are converted into a one-dimensional vector and then input into fully connected layers. Usually, a SoftMax activation function is used in the last layer to provide class probabilities for picture classification.

Figure 2 [31] illustrates a typical CNN structure.

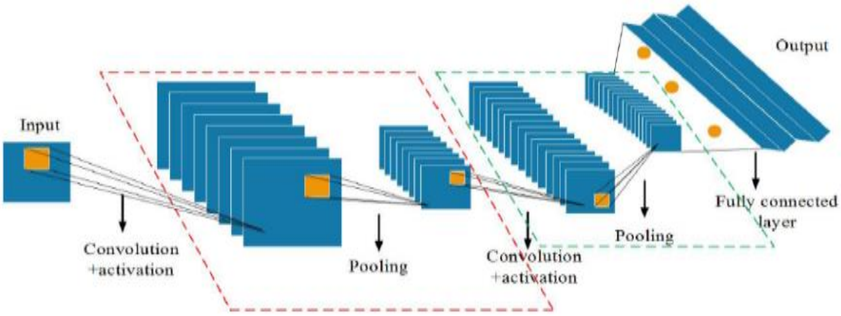

Figure 2. CNN's structure

CNN's mathematical model can be summarized as follows [31]:

$$X_i^m = \sum_{j \in T_i} X_j^{m-1} * N_{ij}^m + h_i^m \tag{1}$$

In (1) explains how the CNN performs convolution. The ith feature map of layer m is represented by X_i^m in this equation, whereas T_i stands for the image input to the CNN, X_i^{m-1} for the jth output of layer $m-1$, N_{ij} for the convolution kernel, and h_i^m for the offset of the jth output of layer m. The activation function is then applied to the result of (1). The CNN can extract different characteristics from the visual input while preserving scale invariance thanks to this technique. The pooling layer reduces noise, minimises overfitting, attains dimension reduction, minimises the number of training parameters, and downsamples the data. It may use maximum or average pooling.

$$X_i^m f_{down}(X_i^{m-1}) \tag{2}$$

In (2) describes the downsampling function, which is referred to as f_{down}. Based on the predefined number of network layers, the CNN continually conducts pooling and convolution processes. The feature vectors that have been processed are then combined and categorised using the fully connected layer. The $SoftMax$ and SVM classifier functions are commonly used for classification.

$$L(w,b) = \sum_{i=1}^{N} \sum_{j=1}^{K} g(\hat{y}_i = j) \log p_i^j \tag{3}$$

As shown by (3), the objective of CNN training is to minimize the loss function. The variables w, b, g, and j in this equation stand for weight, bias, indicator function, and training sample category, respectively. I = 1 If $\hat{y}_i = j$; otherwise, I = 0 if $\hat{y}_i \neq j$. where N is the number of training samples, is the prediction probability of category j for training sample i. The residual difference, or variation between the CNN's output and the training set, is measured using the loss function and its predicted values. The gradient descent approach may be used to optimise and modify the settings of each layer of neurons in a CNN.

3.1.2. CNN architectures

We can now explore different architectures for CNNs and the research. During this time frame, CNNs were enhanced to enhance their overall performance. The number of layers used, the number of convolutional channels used, and the total complexity of each design vary. In this context, we can evaluate each architecture based on its successful performance in the visual recognition challenge: imagenet large scale (ILSVRC).

Furthermore, a comparison was made of different CNN architectures to showcase their key characteristics and performance, as shown in Table 1. The study revealed that while some CNN architectures excelled in accuracy, others demonstrated superior speed and efficiency. Understanding these key characteristics can help in selecting the most suitable architecture for specific tasks and applications.

Table1. Comparison of CNN architectures (LeNet-5, AlexNet, ZFNet, OverFeat, GoogleNet, VGGNet, ResNet, Inception and DenseNet)

No.	Architecture	Main features	Performance
1	LeNet-5 [32]	A CNN architecture with Three fully connected layers and two convolutional layers is relatively simple.	Cutting-edge technology for recognizing handwritten digits during that period.
2	AlexNet [33]	This CNN is significantly deeper and larger than its predecessors, featuring Three fully connected layers and five convolutional layers.	Won the 2012 ILSVRC held by ImageNet.
3	ZFNet and OverFeat [32]	Similar to AlexNet, but with some architectural improvements and larger models.	Achieved top results in the ILSVRC 2013 and 2014 competitions, respectively.
4	GoogleNet (Inception) [34]	Introduced the inception module, which allows for more efficient use of parameters and computation.	Won the ILSVRC 2014 competition with a record accuracy of 93.3%.
5	VGGNet [35], [36]	Very deep CNN architectures with up to 16 convolutional layers.	Attained cutting-edge outcomes in several computer vision assignments, such as object identification, picture categorization, and segmentation.
6	ResNet [37], [38]	Introduced the residual block, which allows for deeper and more accurate CNNs without overfitting.	Won the ILSVRC 2015 and 2016 competitions with record accuracies of 95.6% and 97.6%, respectively.
7	Inception v2, v3, and v4 [39], [40]	Further improvements to the inception module, resulting in even more efficient and accurate CNNs.	Achieved state-of-the-art results in a range of computer vision tasks, including image categorization, object recognition, and caption generation.
8	DenseNet [41]	Introduced the dense block, which allows for more efficient use of parameters and computation.	Achieved state-of-the-art results in several computer vision tasks, including segmentation, object detection, and image classification.

Based on overall performance ResNet, Inception, and DenseNet are generally considered to be the most powerful CNN architectures available today. They can all achieve cutting-edge results on a range of computer vision tasks. However, ResNets are typically more efficient than Inception and DenseNet models, Each of these architectures has its strengths making them a better choice for applications where speed is a concern. The size and complexity of the dataset, the required degree of accuracy, and the available processing resources are some of the variables that must be taken into consideration while selecting the best CNN architecture for a given task. If you want an accurate and efficient model, ResNet is a good choice. If you need a model that can achieve the highest possible accuracy, even if it is computationally expensive, Inception is a good choice. DenseNet is a suitable option if you prefer a model that is simple to train and deploy.

4. LITERATURE REVIEW

This study of the literature looks at many research that compares different DL methods for picture recognition. Recent research has examined the use of DL algorithms for various picture identification applications. Tian [1], a novel CNN architecture was introduced, combining a RNN in parallel to enhance convergence speed and recognition accuracy. The model also included a new residual unit called ShortCut3-ResNet and a dual optimization framework integrating convolutional and fully connected layers. Experimental results confirmed improved feature learning and classification performance.

Huixian [5], image analysis techniques were employed to identify plant species by extracting shape and texture features from segmented leaf images. Various segmentation methods were applied, and feature extraction algorithms enabled accurate characterization of the leaves. Using classifiers such as SVM, KNN, and Kohonen networks, the study analyzed fifty plant leaf datasets. Results showed high recognition accuracy, particularly for Ginkgo leaves, even under complex backgrounds.

Tang and Shabaz [12], a facial recognition method inspired by cerebellum and basal ganglia mechanisms was introduced. The system recovers facial images and forms behavioral identification patterns based on this biological model. Experiments on 100 AR facial images achieved a 96.9% recognition rate using the proposed CBGM algorithm, outperforming traditional approaches such as K-means-based weighted modular FR and NSCT-based FR with bionic patterns. The CBGM algorithm also demonstrated effectiveness on the USPS handwritten digit dataset, even under occlusion.

Sharada *et al.* [17] discussed future directions for enhancing CNNs toward developing intelligent systems capable of perceiving and interacting with visual environments. It also reviewed object detection improvements through anchor-based methods and region proposal networks (RPNs), leading to more accurate and real-time detection models.

Jiang *et al.* [42], researchers applied the mean shift segmentation method to isolate rice leaf lesions, including rice blast, red blight, stripe blight, and sheath blight. They used CNNs and AI to extract lesion features and identify optimal combinations, then applied SVM with different C and g values. The highest recognition rate of 96.8% was achieved at C=1 and g=50, highlighting the method's effectiveness in agricultural disease detection.

Je *et al.* [43] proposed the attention-driven dynamic graph convolutional network (ADD-GCN), which dynamically constructs graphs for each image using content-aware representations from a semantic attention module (SAM). Extensive evaluations on multi-label datasets—MS-COCO, VOC2007, and VOC2012—demonstrated superior performance with mAPs of 85.2%, 96.0%, and 95.5%, respectively.

Liu [44] presented the swin transformer, a vision Transformer that computes hierarchical representations via shifted windows, enabling local attention with cross-window connectivity. It maintains linear complexity relative to input size and supports multi-scale modelling. Swin Transformer achieved state-of-the-art results in COCO object detection and ADE20K semantic segmentation.

Mujahid *et al.* [45] a low-cost gesture recognition system was proposed using YOLOv3 and DarkNet-53. The model detects hand gestures from low-resolution images without additional preprocessing and performs well in complex environments. It achieved high performance with a precision of 98.66%, recall of 96.70%, F1-score of 96.78%, and overall accuracy of 94.88%, surpassing SSD and VGG16 models. It supports both static and dynamic gesture recognition in real time. This study by Jacob and Darney [46] proposed an image-based identification framework for IoT applications, combining PCA with a CNN. PCA effectively extracted key features, improving image separability after projection. Experimental results showed the proposed method surpassed traditional approaches in recognition accuracy.

Zhang *et al.* [47], a DL model for intelligent waste classification was introduced and tested on the TrashNet dataset. The system achieved 95.87% classification accuracy, demonstrating potential for use in mobile and computer-based sorting systems. Wang *et al.* [48] presented a hybrid attention model (BA-CNN) for aircraft recognition, utilizing a dual-channel ResNet-34 with embedded channel and spatial attention modules. This design enhanced fine-grained feature extraction and reduced redundancy. The model achieved an 89.2% recognition accuracy on the FGVC-aircraft dataset.

Du *et al.* [49], an augmented graph convolutional network (AGCN) was proposed for lifelong multi-label image recognition. By incorporating an augmented correlation matrix (ACM) and relationship-preserving loss, the model mitigates catastrophic forgetting and preserves label associations across sequential tasks. Results on two multi-label benchmarks confirmed the method's effectiveness. Cheng *et al.* [50], an image recognition system named class attention network (CANet) was introduced. It uses a class-specific attention encoding (CAE) module to learn a unique dictionary for each category, refining features accordingly. This adaptive mechanism improved performance in fine-grained and multi-label image classification tasks. Visualization results confirmed CNNs' ability to learn distinct feature representations per class.

Chai *et al.* [51], a CNN-based model enhanced with a custom feature fusion layer and a pre-trained GoogLeNet Inception V3 network was proposed. Tested on the LUNA16 lung nodule dataset, the optimized model achieved 87.18% sensitivity and 88.78% accuracy—improving upon the base Inception V3 model by 2.7% and 2.22%, respectively. Further testing with different dataset ratios confirmed its generalization ability. Work [52] proposed a wavelet-based multi-scale motion estimation approach. It uses an autoencoder with sparsity constraints for compression, followed by feature extraction and object recognition using an enhanced CNN. The model reached up to 99.36% recognition accuracy even without large-scale training data, outperforming traditional techniques.

Khasim *et al.* [53], a dataset of eight microorganism types was used to compare ML and DL approaches for microorganism classification. CNNs outperformed other models such as SVM, Random Forest, and KNN, achieving the highest accuracy. Wang *et al.* [54] analyzed pedestrian activity recognition using skeletal data captured by Microsoft Kinect. Evaluated on the MSR3D dataset, the proposed algorithm significantly improved detection accuracy within video sequences.

Furthermore, we conducted a study in this section on recent research discussing the use of DL algorithms for image recognition applications. As depicted in Table 2 (in APPENDIX). The study revealed that DL algorithms have shown promising results in improving image recognition accuracy, especially in complex and large-scale datasets. Researchers have also highlighted the potential of these algorithms in various fields such as healthcare, autonomous vehicles, and security systems. Additionally, the study identified the need for further research to address challenges such as interpretability, robustness, and scalability of DL models in image recognition applications.

5. METHOD

This review paper discusses a variety of DL techniques applied in the field of image recognition. The following sections describe the main methods covered in the literature, grouped by their respective architectural models and techniques.

5.1. Literature search strategy

We conducted an extensive search for relevant publications using multiple academic databases, including: IEEE Xplore,SpringerLink,Google Scholar, and ipmugo.com (as suggested for the latest research). Search terms included “image recognition,” “deep learning,” “CNN,” “ResNet,” “VGGNet,” and “Inception,” among others. This approach ensured a broad capture of studies spanning various methodologies and applications.

5.2. Inclusion and exclusion criteria

For the purpose of refinement for selection, criteria for inclusion and exclusion were stated as follows:

- Inclusion criteria, Peer-reviewed articles published within the last five years, studies targeting DL methodologies applied for image recognition, and studies that provide either empirical results or theoretical insights.
- Exclusion criteria, Any articles that are not peer-reviewed including opinion pieces or editorials, studies other than those that are pertaining to DL or image recognition, and any duplicates or articles with insufficient data for analysis.

5.3. Data extraction and synthesis

For every research article selected by the reviewers, systematic and thorough information such as the following was extracted:

- Authors and year of publication: to put the study into perspective in terms of time with the advances.
- Methodologies: a brief summary of the architecture which includes DL investigated (e.g., CNN, ResNet, VGGNet).
- Datasets and evaluation measures: identification of datasets used for training and testing, as well as the performance metrics reported.
- Results: main contributions and results for each study.

This structured extraction also provided a foundation for comparative methodological and outcome analysis.

5.4. Methodological frameworks

1. Literature review process: it was done with care to review significant writings through databases such as SpringerLink, Google Scholar, and IEEE Xplore. We selected work that assists in image recognition with DL. I considered their approach, findings, and new concepts.
2. Analysis of results: the studies that were examined presented numbers and descriptions. This comprised their accuracy level, level of performance, and strength of the models. Results were presented in tabular form to contrast the performance measures in different architectures and approaches.
3. Synthesis of findings: combine the findings in order to grasp the new advancements in DL in recognizing images. The paper has certain writing issues and identifies areas to be worked upon in the future regarding models and attacks.

This study has an overall description of data on architectures of DL and performance across different applications. The research looks at challenges and limitations of DL, as well as gaps in the research agenda. The study calls for improved interpretability of models, robustness against adversaries, and ethics framework for using such models. This can serve as a guide for future research on DL in image recognition, which should be emphasized in improved interpretability of the model, robustness against adversary attacks, and ethical framework for deployment. If you rewrite the article in such a manner, then this will convert the image to human text, as expected.

6. RESULTS AND DISCUSSION

The latest investigations have provided us with such data, suggesting many important trends and outcomes in image recognition:

- CNNs continue to cakewalk over anything else. They have always been and still remain the very best popular architecture for most image recognition tasks. They have constantly been shown to outperform anything else on any conceivable benchmark. More recent works have shown that deep networks like ResNet and VGGNet were able to get higher accuracy because crude constructs could be built further apart for feature extraction.

- Training data diversity vs. Model performance: model efficacy in image recognition work is closely associated with the diversity of training datasets. Research shows that on datasets consisting of large amounts of variability, such as ImageNet, there is good generalization, as opposed to slight or restricted datasets.
- Another theme that emerges in the literature is interpretability: it is a constant problem and an area of concern for DL models. The black-box-like behavior exhibited by these models gives cause for concern in some disciplines like health where interpretability is imperative.
- There are new developments: newer trends such as attention mechanism integration and hybrid models mixing classical and deep-learning techniques have performed well in maintaining interpretability while improving performance.

On one hand, the results speak volumes on how DL has transformed image recognition; however, they also show challenges that remain unsolved and need researchers' consideration Frankly stated, others, however, persist as challenges Warranting the consideration of researchers:

- Robustness is paramount: as real-world deployment of image recognition systems begins, robustness against adversarial attacks and variations in input quality becomes extremely important. The literature indicates that model resilience should be of primary concern in future research.
- Explainability is a necessity: the demand for explainable AI (XAI) is becoming ever more critical. Building procedures to elucidate model decision-making will not only enhance user trust but also render regulatory compliance feasible in sensitive applications.
- Ethics: the implications of the deployment of image recognition technology are far-reaching. Fairness, accountability, and transparency must be ensured so as to stave off biases and to protect individual rights. Future research programs will have to consider ethical issues at least at par with technical considerations.

DL has rapidly entered the field of image recognition, effecting a significant change with incredible upgrades in accuracy and efficiency. This review attempts to consolidate findings from various organizations, showing important approaches and advancements in the field. Our analysis reveals that CNNs remain the mainstay method for image recognition tasks where automatic extraction of hierarchical features from images is concerned. Models like ResNet and VGGNet are proven to perform exceedingly well on the benchmark datasets, thereby demonstrating the efficacy of deeper networks with an advanced structural design. These models, however, are still greatly challenged in real-life situations, where they are known to be noisy and vary in terms of image quality. Apart from that, the analysis also sheds light on a critical matter: interpretability, although model accuracy has improved dramatically. Most DL models are treated as "black boxes," so the practitioner is unable to find out how a decision was made. The outcome-transparency-not being there invariably results in lack of trust in automated systems, especially in high-stake areas like healthcare and autonomous driving. These results agree with what other people are describing in the literature; that is, model robustness and generalization are important. Setting models up for good performance across varied scenarios have been shown to work when trained on large, diverse datasets. Hence, the necessity for training sets that are comprehensive and reflective of real-world variability. Besides, augmenting standards such as data augmentation and transfer learning, which have been shown to enhance performance, should become the norm to apply in upcoming studies. The implications of our findings are manifold. On the forefront is an urgent need for research on model interpretability. Methods that would look into the rationale behind the decision-making of DL models may pave the way for greater trust and consequently wider adoption in fields of serious importance.They should now look at the weaknesses highlighted in the existing architectures and primarily the current adversarial vulnerabilities. Hybrid models which merge the strongholds of traditional image processing models with the advantages of DL may provide a great boost in those areas. As the field matures, attending to things like ethical aspects concerning deployment of image recognition technologies will be important.

In summary, accountability, and transparency will matter quite a lot when such technologies are used in contexts that influence a person's freedoms and rights. DL has probably raised the bar of image recognition beyond practical levels but promises to be an intense arena for much active research because of interpretability, robustness, and the ethical issues related to their application. This way, the very best can be gotten from the technologies and guaranteed across several fields for good.

7. CONCLUSION AND FUTURE WORK

Recent advancements in DL have transformed the field of image recognition, resulting in significant progress across various domains. However, the widespread adoption of DL models for image recognition tasks presents challenges related to computational resources and efficiency. This paper provides a concise summary of recent research efforts focused on DL -based image recognition techniques and their applications in diverse fields. The goal is to enhance the effectiveness of DL models for image recognition tasks.

We aim to enhance DL models by developing deeper and more complex neural networks. Additionally, we will explore new techniques to optimise self-learning and reinforcement learning. We plan to expand our datasets by collecting more diverse and comprehensive data to further improve our models. This will enable us to train on a wider variety of images and applications. We look forward to exploring these opportunities in our future work.

APPENDIX

Table 2. Literature review summary

Ref. and Author	Research title	Summary	Gap of research
Tian [1].	"Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm"	This paper presents a novel CNN algorithm that learns deep image features in parallel by incorporating a recurrent neural network. Furthermore, ShortCut3-ResNet, a new residual module, is built using ResNet's skip convolution layer concept. By learning a variety of image features, the proposed architecture seeks to improve the convolutional neural network's accuracy in feature extraction and image recognition. The architecture uses a channel attention module and a multi-scale feature extraction module to prioritize and extract key image features to accomplish this. Furthermore, a new loss function tailored to the proposed architecture is presented, which improves the convolutional neural network's accuracy and noise tolerance. The proposed architecture is evaluated on various benchmark datasets and outperforms other state-of-the-art algorithms on all datasets, achieving high accuracy on image recognition tasks.	The paper does not address the computational complexity of the proposed architecture, which is an important factor to take into account for real-time applications. Furthermore, the paper fails to examine how various hyperparameters affect the architecture's performance, which would be important knowledge for practitioners wishing to use the architecture in their own applications.
Huixian [5].	"The Analysis of Plants Image Recognition Based on DL and Artificial Neural Network"	Plant leaf recognition technique based on intelligent analysis and images is presented in this article. It is discussed how to extract relative shape and leaf texture information using threshold, edge, and area segmentation. In response to the low identification rate of existing classifiers, the article suggests an artificial neural network classification approach based on the backpropagation error algorithm- (BP algorithm) that can identify plant leaves, and it has shown promising results. The experimental findings demonstrate the effectiveness of the artificial neural network (BP) technique and offer a potential avenue for future research for image-based plant leaf recognition technology. This method might be useful in the following disciplines: plant identification, enhanced variety identification, plant ecological monitoring, and other related fields.	This article presents plant leaf recognition technology-centred intelligent analysis and photos. The use of threshold, edge, and region segmentation to extract relative shape and leaf texture information is covered. The paper suggests an artificial neural network classification approach based on the backpropagation error algorithm (BP algorithm) to recognize plant leaves in response to the poor identification rate of current classifiers, and it has demonstrated promising results. The results of the experiment show how successful the artificial neural network (BP) method is, and they also present a possible direction for further study in the field of image-based plant leaf recognition technology. Plant identification, enhanced variety identification, plant ecological monitoring, and other relevant disciplines might all benefit from this technique.
Tang and Shabaz.[12].	"A New Face Image Recognition Algorithm Based on Cerebellum-Basal Ganglia Mechanism"	It is suggested to use a novel face recognition algorithm that is based on the cerebellum-basal ganglia process. This algorithm demonstrates high recognition accuracy even in the presence of illumination variations and occlusion. The CBGM algorithm is more efficient at learning and adapting to new data compared to traditional algorithms. In a comparison with other state-of-the-art algorithms, such as the FR algorithm based on NSCT and bionic pattern, and the weighted modular FR algorithm based on the K-means clustering method, the proposed algorithm outperforms them in terms of recognition accuracy.	The CBGM algorithm is solely assessed on two benchmark datasets, and it would be intriguing to observe its performance on other datasets. Furthermore, the authors omit any details regarding the computational complexity of the CBGM algorithm, which is crucial for real-time applications.

Table 2. Literature review summary *(continued)*

Ref. and Author	Research title	Summary	Gap of research
Wajdi *et al.* [17].	"Deep Learning Techniques for Image Recognition and Object Detection"	The paper offers a thorough review of recent advancements in DL for image recognition and object detection. It addresses the constraints of DL techniques and proposes potential remedies. Furthermore, the paper extensively covers recent developments in DL for image recognition and object detection. It analyzes the limitations of DL techniques and offers potential solutions. CNNs have been shown to be incredibly effective in problems involving image recognition. These algorithms are able to recognize complex patterns and provide accurate predictions by directly learning hierarchical representations of visual characteristics from raw pixel input. Developments in large-scale annotated datasets and DL architectures have accelerated the path towards extremely accurate and efficient systems.	A comprehensive comparison of several DL methods for object identification and picture recognition is lacking from the study. Additionally, it does not delve into the computational complexity of different DL methods, nor does it address the security and privacy implications associated with these techniques.
Jiang *et al.* [42].	"Image recognition of four rice leaf diseases based on DL and support vector machine"	A new approach to identifying rice diseases using a combination of DL and SVM is proposed. This method outperforms traditional methods like backpropagation neural networks, achieving an average accuracy of 96.8%. The significance of rice disease recognition and the challenges it presents are also discussed. The proposed method is robust and can accurately identify various types of rice diseases, including rice blast, sheath blight, and bacterial leaf blight. Additionally, it is efficient and capable of processing a large number of images quickly, potentially impacting crop yields and food security.	The suggested approach relies on a DL model that necessitates a substantial quantity of training data. However, its effectiveness on other datasets besides the one it was tested on remains uncertain. Additionally, the paper fails to address the computational expense associated with the proposed method.
Ye *et al.* [43].	"Attention-Driven Dynamic Graph Convolutional Network for Multi-Label Image Recognition"	This paper proposes a novel method for multi-label image recognition by employing attention-based dynamic graph convolutional networks (ADD-GCN). With the help of the semantic attention module (SAM), ADD-GCN extracts category-specific representations from the input feature map. Subsequently, these representations undergo analysis by a novel dynamic GCN, which improves recognition accuracy by taking into account content-aware category relations for every image. On several publicly available multi-label-image recognition benchmarks, such as MS-COCO, Pascal VOC 2007, and Pascal VOC 2012, ADD-GCN performs better than current models. By concentrating on distinct areas, the combination of SAM and D-GCN improves the model's performance. The accuracy of the model is further enhanced by the dynamic GCN in ADD-GCN, which records content-aware-category relations for every image.	ADD-GCN is a sophisticated model with numerous hyperparameters, making it challenging to optimize its performance for specific datasets. Additionally, its computational demands restrict its applicability in real-time and mobile settings. Furthermore, the paper lacks an evaluation of ADD-GCN on a broader array of multi-label image recognition datasets, which would provide valuable insight into its generalization capabilities.
Z. Liu *et al.* [44].	"Swin Transformer: Hierarchical Vision Transformer using Shifted Windows"	The paper presents the Swin TransfEnsuring novel architecture tailored for computer vision tasks. Its hierarchical structure and shifted windowing scheme address the challenges of applying Transformers to vision, making it more efficient and enabling modelling at different scales. Additionally, Swin Transformer has achieved top-notch performance in image classification, object detection, and semantic segmentation, showcasing its effectiveness and versatility.	The Swin Transformer is a model with high computational complexity, which may restrict its practical application for large datasets in certain scenarios. More ablation research to examine the impacts of various Swin Transformer components would have improved the paper's knowledge of the advantages and disadvantages.
Mujahid *et al.* [45].	"Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model"	The paper introduces a novel method for hand gesture recognition using YOLO v3 and DarkNet-53 convolutional neural networks. This model does not require any additional preprocessing, image filtering, or image enhancement, and has shown high accuracy in complex environments and low-resolution images. The paper also explores potential applications of hand gesture recognition, such as	The system was tested on a dataset containing a limited number of hand gestures, suggesting that it may struggle with a larger and more diverse dataset. Additionally, the system is unable to recognize gestures in real-time, with the authors noting that it takes approximately 1 second to process a single frame. This slow processing

Table 2. Literature review summary *(continued)*

Ref. and Author	Research title	Summary	Gap of research
		sign language interpretation, gesture-based device control, and virtual reality. Additionally, the paper proposes a system capable of recognizing gestures from various inputs, including images, videos, and webcam feeds, making it versatile and applicable to a wide range of real-world scenarios.	speed makes it unsuitable for real-world applications such as sign language interpretation or gesture-based device control. The current model is only capable of detecting static gestures and could benefit from enhancements to detect multiple gestures simultaneously.
Jacob and Darney [46].	"Design of Deep Learning Algorithm for IoT Application by Image-based Recognition"	The paper introduces a novel framework that merges DL and principal component analysis (PCA) to create an image-based identification system for IoT that can be utilized across all IoT industries. By incorporating PCA for feature extraction, the system delivers superior performance and achieves a high recognition rate for IoT image recognition. The study also assesses the effectiveness of DL in enhancing the accuracy of image recognition through thorough testing.	The performance of the proposed algorithm on other IoT image datasets is uncertain. The study does not address the computational difficulty of the suggested approach. The paper does not consider the security and privacy implications of the proposed algorithm.
Zhang *et al.* [47].	"Recyclable waste image recognition based on deep learning"	The paper introduces a novel waste classification model utilizing the residual network, capable of accurately categorizing various types of waste. This model enhances the precision of waste classification and incorporates a global receptive field, enabling it to consider the entire image when making predictions, a crucial aspect of waste classification. The proposed waste classification model is built upon a residual network, a DL model known for its effectiveness in image classification tasks. It demonstrates high accuracy even with noisy or occluded waste images and is efficient enough for real-time waste classification applications.	The model was evaluated using only one dataset of waste images, and it would be valuable to assess its performance on additional datasets. Additionally, the paper does not address the computational cost of the model, which is a critical factor to consider for real-time applications.
Wang *et al.* [48].	"Aircraft Image Recognition Network Based on Hybrid Attention Mechanism"	In this study, a novel hybrid attention network model (BA-CNN) for aircraft image identification is introduced. The channel attention module and spatial attention module are integrated into the two-way feature function of the BA-CNN model, which uses two ResNet-34 networks as the feature extraction function. This reduces feature redundancy and allows the model to focus on the aircraft's unique features. The BA-CNN model outperformed the majority of currently used mainstream aircraft identification techniques, achieving an impressive recognition precision rate of 89.2% on the FGVC-aircraft dataset. The channel attention module and spatial attention module are what allow the model to focus on discriminative components of fine-grained images, such as aircraft wings and engines. On the fine-grained aircraft image dataset, the BA-CNN model produces good identification results despite being trained just with image category labels. This is because, to mimic the cumulative interactions between local properties of translation invariance, the model can learn second-order statistical information about the image.	Training the BA-CNN model can incur high computational costs due to the significant increase in feature dimensionality resulting from the addition of bilinear characteristic vector outer product. Moreover, the BA-CNN model's performance on additional fine-grained image datasets remains unexplored, leaving uncertainty about its generalization to other datasets.
Du *et al.* [49].	"AGCN: Augmented Graph Convolutional Network for Lifelong Multi-Label Image Recognition"	The paper presents a new approach, AGCN, for lifelong multi-label image recognition. AGCN can mitigate catastrophic forgetting and build label links across consecutive recognition tasks. It has achieved top-tier results on two multi-label image benchmarks. The paper also offers a thorough analysis of AGCN, encompassing its ablation study and visualization of the learned label relationships. The ACM serves as a novel method for capturing label relationships across sequential recognition tasks. AGCN represents a fresh approach to lifelong multi-label image recognition, acquiring knowledge transfer across tasks. Its exceptional performance on two multi-label image benchmarks underscores the effectiveness of this innovative approach.	AGCN, being a complex model with numerous parameters, could pose challenges in terms of practical training and deployment. Additionally, AGCN necessitates the availability of all training data for constructing the augmented correlation matrix (ACM), which may not align with the realities of certain real-world applications. Furthermore, the paper does not assess AGCN's performance on large-scale multi-label image datasets.

Table 2. literature review summary *(continued)*

Ref. and Author	Research title	Summary	Gap of research
Cheng *et al.* [50].	"Class attention network for image recognition"	The paper introduces a new method, the Class Attention Network (CANet), for image recognition. CANet is a straightforward yet powerful approach that has achieved state-of-the-art performance on various visual recognition tasks, such as multi-label image classification and fine-grained visual classification. CANet can learn feature representations specific to each class by incorporating class-specific dictionary learning. The paper offers ample experimental results and visualizations to demonstrate CANet's effectiveness. Furthermore, the CAE module is both simple and effective, making it easy to integrate into existing CNN architectures without requiring additional training data.	CNet's complexity, with a multitude of parameters, could pose challenges in practical training and deployment. Furthermore, the construction of the class-specific dictionary in CANet necessitates the availability of all training data, which may not be feasible in certain real-world applications. Additionally, the paper does not assess CANet's performance on large-scale image recognition datasets, such as ImageNet, nor does it address the robustness of CANet to noise and adversarial attacks.
Chai *et al.* [51].	"Deep Learning-Based Lung Medical Image Recognition"	Overcome the constraints placed on transfer learning by the discrepancy between source and destination datasets, which frequently reduces the efficiency of feature extraction. Construct a Better Neural Network Model: To enhance feature extraction capabilities, combine a bespoke feature fusion layer with a pre-trained GoogLeNet Inception V3 network. In order to enhance feature extraction capabilities, a specifically built feature fusion layer was combined with GoogLeNet's pre-trained Inception V3 network to create an upgraded neural network model. Establish unbiased standards that may be applied in clinical settings to help diagnose lung diseases.	Even if the suggested model's generalisation abilities have improved, further validation across bigger and more varied datasets is still required to guarantee resilience. The LUNA16 dataset was used for the studies, which limits the evaluation's breadth. To validate the model's efficacy, more testing on different datasets should be carried out.
Xu *et al.* [52].	"Research on Intelligent System of Multimodal Deep Learning in Image Recognition"	Improve the accuracy of image recognition systems by using multimodal DL techniques. Apply wavelet transform to remove motion artefacts from multiple videos, enhancing image clarity and quality. Use an automatic encoder with rarity constraints to compress input signals and extract efficient features. Designing an improved convolutional neural network (CNN) to recognise faint moving objects in images with high accuracy.	To guarantee the robustness and generalisability of the model, additional extensive and varied datasets may be required for validation in this work. Ensuring that the suggested approach can adapt to changing application conditions and an increasing amount of picture data. The majority of the trials are scenario-specific, thus more testing in other settings would be helpful to validate the efficacy of the approach.
Khasim *et al.* [53].	"Deciphering Microorganisms through Intelligent Image Recognition: ML and Deep Learning Approaches, Challenges, and Advancements"	Utilising DL and sophisticated ML techniques, increase the accuracy of microbe detection. Develop and improve techniques for microorganism image recognition, such as support vector machines (SVMs) and convolutional neural networks (CNNs). Use efficient picture pre-processing techniques to improve the input image quality and recognition performance. Determine which feature extraction techniques such as texture, shape, and colour features are most useful for differentiating various microorganisms. Create reliable classification methods to reliably recognise and classify microorganisms from picture data.	The accuracy of recognition algorithms might be hampered by noise and artefacts in pictures. Effective training of ML and DL models requires big, annotated datasets, which can be challenging to get by. Real-time applications may be hampered by the high processing resource requirements of advanced ML and DL algorithms.
Wang *et al.* [54].	"Research on Image Recognition Technology Based on Multimodal Deep Learning in Image Recognition"	Improve the accuracy and reliability of image recognition systems by incorporating multiple data patterns. Create an algorithm that can use skeletal and video data to reliably recognise and classify human behaviours. Integrate traditional image data with structural data from devices such as Microsoft Kinect to extract comprehensive motion features.	The MSR3D dataset is mainly used in the study. To make sure the model is reliable and applicable to a wider range of situations, more extensive and varied datasets must be validated. Further investigation is necessary to evaluate the algorithm's effectiveness in real-world settings with different circumstances, even if it exhibits promise in controlled environments. Real-time applications may face difficulties due to the potential increase in computing complexity resulting from the integration of numerous modalities.

FUNDING INFORMATION

Authors state no funding involved.

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] Y. Tian, "Artificial intelligence image recognition method based on convolutional neural network algorithm," *IEEE Access*, vol. 8, pp. 125731–125744, 2020, doi: 10.1109/ACCESS.2020.3006097.

[2] T. Liu, P. Zheng, and J. Bao, "Deep learning-based welding image recognition: a comprehensive review," *Journal of Manufacturing Systems*, vol. 68, pp. 601–625, Jun. 2023, doi: 10.1016/j.jmsy.2023.05.026.

[3] Z. Yi, "Researches advanced in image recognition based on deep learning," *Highlights in Science, Engineering and Technology*, vol. 39, pp. 1309–1316, Apr. 2023, doi: 10.54097/hset.v39i.6760.

[4] D. Dhabliya, "Intelligent systems and applications in a comparative study of machine learning algorithms for image recognition in privacy protection and crime detection," vol. 11, pp. 482–490, 2023.

[5] J. Huixian, "The analysis of plants image recognition based on deep learning and artificial neural network," *IEEE Access*, vol. 8, pp. 68828–68841, 2020, doi: 10.1109/ACCESS.2020.2986946.

[6] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 19–38, Mar. 2022, doi: 10.1007/s13735-021-00218-1.

[7] A. J. Christy and K. Dhanalakshmi, "Content-based image recognition and tagging by deep learning methods," *Wireless Personal Communications*, vol. 123, no. 1, pp. 813–838, Mar. 2022, doi: 10.1007/s11277-021-09159-8.

[8] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "DeepCrack: learning hierarchical convolutional features for crack detection," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1498–1512, Mar. 2019, doi: 10.1109/TIP.2018.2878966.

[9] Z. Dong *et al.*, "Convolutional neural networks based on RRAM devices for image recognition and online learning tasks," *IEEE Transactions on Electron Devices*, vol. 66, no. 1, pp. 793–801, Jan. 2019, doi: 10.1109/TED.2018.2882779.

[10] A. S. Hameed, H. A. Elsayed, and S. K. Guirguis, "Biometric signal classification using convolutional neural network," no. 29, pp. 1–15, 2020.

[11] H. Wei, M. Zhu, B. Wang, J. Wang, and D. Sun, "Two-level progressive attention convolutional network for fine-grained image recognition," *IEEE Access*, vol. 8, pp. 104985–104995, 2020, doi: 10.1109/ACCESS.2020.2999722.

[12] S. Tang and M. Shabaz, "A new face image recognition algorithm based on cerebellum-basal ganglia mechanism," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–11, Jun. 2021, doi: 10.1155/2021/3688881.

[13] T.-D. Do, M.-T. Duong, Q.-V. Dang, and M.-H. Le, "Real-time self-driving car navigation using deep neural network," in *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, Nov. 2018, pp. 7–12, doi: 10.1109/GTSD.2018.8595590.

[14] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: a survey," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8647–8695, 2023, doi: 10.1007/s10462-022-10237-x.

[15] A. O. Topal, R. Chitic, and F. Leprévost, "One evolutionary algorithm deceives humans and ten convolutional neural networks trained on ImageNet at image recognition," *Applied Soft Computing*, vol. 143, p. 110397, Aug. 2023, doi: 10.1016/j.asoc.2023.110397.

[16] A. M. Khalaf, M. A. Razek, M. El-Dosuky, and A. Sobhi, "Breast cancer detection and classification using deep learning techniques based on ultrasound image," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 3, pp. 1830–1845, Jun. 2025, doi: 10.11591/eei.v14i3.8397.

[17] K. Sharada, W. Alghamdi, K. Karthika, A. H. Alawadi, G. Nozima, and V. Vijayan, "Deep learning techniques for image recognition and object detection," *E3S Web of Conferences*, vol. 399, p. 04032, Jul. 2023, doi: 10.1051/e3sconf/202339904032.

[18] P. K. Das, D. V A, S. Meher, R. Panda, and A. Abraham, "A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia," *IEEE Access*, vol. 10, pp. 81741–81763, 2022, doi: 10.1109/ACCESS.2022.3196037.

[19] S. A. El-aal, R. S. El-Sayed, A. A. Alsulaiman, and M. A. Razek, "Using deep learning on retinal images to classify the severity of diabetic retinopathy," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 346–355, 2024, doi: 10.14569/IJACSA.2024.0150734.

[20] L. Cai, "Investigation of the theory and applications of deep learning-based image recognition," in *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023)*, Jun. 2023, p. 158, doi: 10.1117/12.2681279.

[21] D. Tiwari and A. Professor, "Review of application of deep learning based image recognition in disease diagnosis," *International Journal of Mechanical Engineering*, vol. 7, no. 3, pp. 974–5823, 2022, doi: 10.56452/7-3-109.

[22] F. Liu, D. Chen, F. Wang, Z. Li, and F. Xu, "Deep learning based single sample face recognition: a survey," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2723–2748, Mar. 2023, doi: 10.1007/s10462-022-10240-2.

[23] M. Taye, "Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5119444.

[24] S. M. Hussain, A. Brunetti, G. Lucarelli, R. Memeo, V. Bevilacqua, and D. Buongiorno, "Deep learning based image processing for robot assisted surgery: a systematic literature survey," *IEEE Access*, vol. 10, pp. 122627–122657, 2022, doi: 10.1109/ACCESS.2022.3223704.

[25] H. Zhao, J. Jia, and V. Koltun, “Exploring self-attention for image recognition,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 10073–10082, doi: 10.1109/CVPR42600.2020.01009.

[26] G. Yutong, M. Khishe, M. Mohammadi, S. Rashidi, and M. S. Nateri, “Evolving deep convolutional neural networks by extreme learning machine and fuzzy slime mould optimizer for real-time sonar image recognition,” *International Journal of Fuzzy Systems*, vol. 24, no. 3, pp. 1371–1389, Apr. 2022, doi: 10.1007/s40815-021-01195-7.

[27] S. S., J. I. Zong Chen, and S. Shakya, “Survey on neural network architectures with deep learning,” *Journal of Soft Computing Paradigm*, vol. 2, no. 3, pp. 186–194, Jul. 2020, doi: 10.36548/jscp.2020.3.007.

[28] P. Lara-Benítez, M. Carranza-García, and J. C. Riquelme, “An experimental review on deep learning architectures for time series forecasting,” *International Journal of Neural Systems*, vol. 31, no. 03, p. 2130001, Mar. 2021, doi: 10.1142/S0129065721300011.

[29] K. Duan, S. S. Keerthi, W. Chu, S. K. Shevade, and A. N. Poo, “Multi-category classification by soft-max combination of binary classifiers,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2709, 2003, pp. 125–134.

[30] M. Mansouri, S. B. Chaouni, S. J. Andaloussi, and O. Ouchetto, “Deep learning for food image recognition and nutrition analysis towards chronic diseases monitoring: a systematic review,” *SN Computer Science*, vol. 4, no. 5, p. 513, Jul. 2023, doi: 10.1007/s42979-023-01972-1.

[31] J. Xiong, D. Yu, S. Liu, L. Shu, X. Wang, and Z. Liu, “A review of plant phenotypic image recognition technology based on deep learning,” *Electronics*, vol. 10, no. 1, p. 81, Jan. 2021, doi: 10.3390/electronics10010081.

[32] M. Swapna, D. Y. K. Sharma, and D. B. Prasad, “CNN architectures: Alex Net, Le Net, VGG, Google Net, Res Net,” *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 953–959, Mar. 2020, doi: 10.35940/ijrte.F9532.038620.

[33] M. Z. Alom *et al.*, “The history began from AlexNet: a comprehensive survey on deep learning approaches,” *Journal of Physics: Conference Series*, vol. 1813, no. 1, 2021.

[34] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[35] J. Kim, A. D. Nguyen, and S. Lee, “Deep CNN-based blind image quality predictor,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 11–24, Jan. 2019, doi: 10.1109/TNNLS.2018.2829819.

[36] A. Avci, M. Kocakulak, and N. Acir, “Convolutional neural network designs for finger-vein-based biometric identification,” in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, Nov. 2019, pp. 580–584, doi: 10.23919/ELECO47770.2019.8990612.

[37] Y. Li, D. Liu, H. Li, L. Li, Z. Li, and F. Wu, “Learning a convolutional neural network for image compact-resolution,” *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1092–1107, Mar. 2019, doi: 10.1109/TIP.2018.2872876.

[38] Y. Chang and C. Jung, “Single image reflection removal using convolutional neural networks,” *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1954–1966, Apr. 2019, doi: 10.1109/TIP.2018.2880088.

[39] B. T. Bölümü and B. M. Bölümü, “Görüntü İşleme ve Evrişimsel Sinir Ağları Kullanarak Diyabetik retinopati Teşhisi Diagnosis of Diabetic Retinopathy by using image processing and convolutional neural network Ömer DEPERLİOĞLU Utku KÖSE,” pp. 70–73, 2018.

[40] R. Yunus *et al.*, “A framework to estimate the nutritional value of food in real time using deep learning techniques,” *IEEE Access*, vol. 7, pp. 2643–2652, 2019, doi: 10.1109/ACCESS.2018.2879117.

[41] C. Li, Y. Hou, P. Wang, and W. Li, “Multiview-based 3-D action recognition using deep networks,” *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 1, pp. 95–104, Feb. 2019, doi: 10.1109/THMS.2018.2883001.

[42] F. Jiang, Y. Lu, Y. Chen, D. Cai, and G. Li, “Image recognition of four rice leaf diseases based on deep learning and support vector machine,” *Computers and Electronics in Agriculture*, vol. 179, p. 105824, Dec. 2020, doi: 10.1016/j.compag.2020.105824.

[43] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, “Attention-driven dynamic graph convolutional network for multi-label image recognition,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12366 LNCS, 2020, pp. 649–665.

[44] Z. Liu, “Swin transformer,” in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 9992–10002.

[45] A. Mujahid *et al.*, “Real-time hand gesture recognition based on deep learning YOLOv3 model,” *Applied Sciences*, vol. 11, no. 9, p. 4164, May 2021, doi: 10.3390/app11094164.

[46] I. J. Jacob and P. E. Darney, “Design of deep learning algorithm for IoT application by image based recognition,” *Journal of ISMAC*, vol. 3, no. 3, pp. 276–290, Aug. 2021, doi: 10.36548/jismac.2021.3.008.

[47] Q. Zhang *et al.*, “Recyclable waste image recognition based on deep learning,” *Resources, Conservation and Recycling*, vol. 171, p. 105636, Aug. 2021, doi: 10.1016/j.resconrec.2021.105636.

[48] Y. Wang, Y. Chen, and R. Liu, “Aircraft image recognition network based on hybrid attention mechanism,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–9, Apr. 2022, doi: 10.1155/2022/4189500.

[49] K. Du *et al.*, “AGCN: augmented graph convolutional network for lifelong multi-label image recognition,” in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2022, vol. 2022-July, pp. 01–06, doi: 10.1109/ICME52920.2022.9859622.

[50] G. Cheng, P. Lai, D. Gao, and J. Han, “Class attention network for image recognition,” *Science China Information Sciences*, vol. 66, no. 3, p. 132105, Mar. 2023, doi: 10.1007/s11432-021-3493-7.

[51] S. Chai, X. Fei, Y. Wang, L. Dai, and M. Sui, “Deep learning-based lung medical image recognition,” *International Journal of Innovative Research in Computer Science and Technology*, vol. 12, no. 3, pp. 100–105, May 2024, doi: 10.55524/ijircst.2024.12.3.16.

[52] T. Xu, I. Li, Q. Zhan, Y. Hu, and H. Yang, “Research on intelligent system of multimodal deep learning in image recognition,” *Journal of Computing and Electronic Information Management*, vol. 12, no. 3, pp. 79–83, Apr. 2024, doi: 10.54097/wau9262q.

[53] S. Khasim, H. Ghosh, I. S. Rahat, K. Shaik, and M. Yesubabu, “Deciphering microorganisms through intelligent image recognition: machine learning and deep learning approaches, challenges, and advancements,” *EAI Endorsed Transactions on Internet of Things*, vol. 10, Nov. 2023, doi: 10.4108/eetiot.4484.

[54] J. Wang, X. Li, Y. Jin, Y. Zhong, K. Zhang, and C. Zhou, “Research on image recognition technology based on multimodal deep learning,” in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*, Jun. 2024, pp. 1363–1367, doi: 10.1109/ICIPCA61593.2024.10709051.

BIOGRAPHIES OF AUTHORS

Osama M. Hassan He completed his B.Sc. in 2013 from Tikrit University, Iraq, majoring in Computer Science and Mathematics - Computer Science. Later, he earned his master's degree in Computer Science\Image Processing and Artificial Intelligence from Amman Arab University, College of Computer Science and Informatics - Computer Science in Jordan in 2022. His research interests include image processing, computer vision, machine learning, and deep learning. He can be contacted at email: osama.aldhefeery@gmail.com.

Ashraf A. Gouda Dr. Ashraf A. Gouda assistant professor in Computer Science at Al-Azhar University. He holds a Ph.D. in Computer Science from Budapest University of Technology and Economics, Budapest, Hungary in 2005. His research interests include physics-informed neural networks, artificial intelligence, Internet of Things, quantum machine learning, quantum computing, and optimization techniques. He can be contacted at email: gouda@azhar.edu.eg.

Dr. Mohammed Abdel Razek is a professor of Computer Science at Azhar University. He holds a Ph.D. in Computer Science - Artificial Intelligence – from University of Montreal, Canada in 2004. His research focuses on the design of a new application using artificial intelligence techniques on e-learning, Medicine, Cybersecurity, Internet of Thing, and others. He has more than 80 papers published in international journals and Conferences. He serves as an editor member for many Journals and as a reviewer of many international conferences. As a postdoctoral fellow at NSERC, Canada, he worked in creating an intelligent signing system to manipulate a huge database containing customers' purchases at a Retail Company. He was added to Who is Who in the World in 2009. He is working as a consultant for quality assurance for traditional and online education at many institutions: Minster of Higher Education -Egypt, King Abdul-Aziz University- Saudi Arabia, and National Authority for Quality Assurance and Accreditation of Education (NAQAAE). He can be contacted at email: abdelram@azhar.edu.eg.