

Deep Learning-Based Image Classification for COVID-19 Disease: A Comprehensive Review

Sagardeep Ghosh ^{1,✉}, Yogendra Chhetri ², Kakali Das ¹, Amrita Mitra ¹, Pijush Dutta ¹, Pranay Adak ³

1. Department of Computer Science and Engineering, Greater Kolkata College of Engineering and Management, Kolkata, IND

2. Centre for Continuing Education, Indian Institute of Science, Bengaluru, IND

3. Department of Information Technology, Government College of Engineering and Textile Technology, Serampore, IND

Received: May 12, 2026 | Review began: June 11, 2026 | Review ended: August 25, 2026 | Published: September 21, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

The diversity of biological data being generated today is increasing. Among all the data being generated, the coronavirus disease (COVID-19) outbreak has underscored the importance of rapid, reliable diagnostic techniques. In this context, automatic disease diagnosis using deep learning algorithms with medical image data has become an area of research interest. To identify gaps in the literature, investigate methodological trends, and suggest directions for future research, this review paper critically evaluates the use of deep learning techniques in the categorisation of medical images for the diagnosis of COVID-19. It utilised both a systematic bibliometric review and a systematic review of the literature as a methodology to assess the relevant literature. The literature search yielded a total of 1,138 records; 775 articles were then screened, and 46 studies were included in the qualitative analysis of the literature. The increase in research on medical imaging-based diagnosis has been considerable between the years 2020 and 2025, as evidenced by the reviewed studies included in this article. In general, deep learning methodologies yield superior performance when compared to traditional machine learning methods. The studies reviewed demonstrate that the use of convolutional neural network models, transfer learning, ensemble learning, and hybrid approaches yields a high level of performance for classifying chest X-ray and computed tomography images. The review identifies existing research gaps and also indicates future directions of research to create reliable artificial intelligence-based diagnostic tools for use in medical imaging.

Categories: Advances in Biomedical data analysis, Image Processing and Analysis, Deep Learning

Keywords: deep learning, medical image classification, covid-19, chest x-ray, systematic review

Introduction And Background

The COVID-19 pandemic has brought to light an all too common dilemma in modern healthcare: rapid screening comes with inherent volatility, while diagnostic precision is achieved at the expense of speed. Even in settings with limited resources, reverse transcriptase polymerase chain reaction (RT-PCR)-based testing could not scale quickly enough to keep up with growing numbers, shortages in the supply chain, and turnaround times. However, imaging of the chest, specifically chest X-ray (CXR) and computed tomography (CT) scans, was readily available and sensitive enough to detect changes in the lungs caused by infection, even before severe manifestations [1].

How to cite this article:

Ghosh S, Chhetri Y, Das K, et al. (September 21, 2026) Deep Learning-Based Image Classification for COVID-19 Disease: A Comprehensive Review. Cureus J Comput Sci 3 : es44389-026-00286-6. DOI <https://doi.org/10.7759/s44389-026-00286-6>

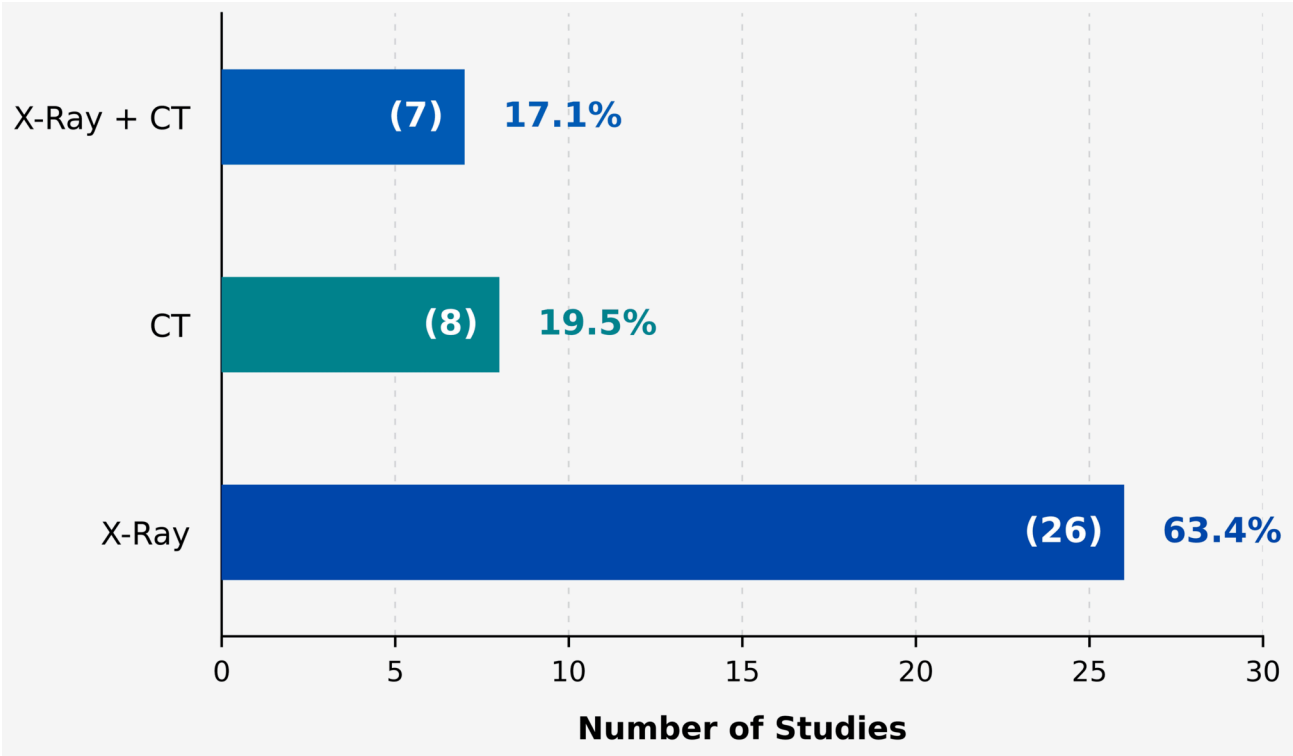


FIGURE 1: Imaging Modalities Used in COVID-19 DL Studies

DL, Deep Learning; COVID-19, Coronavirus Disease; CT, Computed Tomography

As shown in Figure 1, the number of studies based on CXR is much higher than those based on CT and multimodal studies in the current literature due to the fact that X-ray imaging is readily available and deployable during a pandemic. However, the interpretation of imaging investigations was not scalable. During public health emergencies, radiological capacity, distribution, and integrity were and may always be limited. This diagnostic gap was filled by deep learning (DL), particularly convolutional neural networks (CNNs) and their variations, which functioned as an enabling tool rather than an end-to-end replacement [2]. DL was able to rapidly analyse thousands of images, identify anomalous regions, and provide probability guesses for diagnoses in a matter of seconds. DL architectures are able to automatically identify these representation hierarchies from the data itself, in contrast to conventional rule-based image processing chains [3]. They enable DL structures to reveal minute spatial and textural characteristics that human observers might overlook. With hundreds of research studies proposing DL architectures, datasets, and performance evaluation techniques in an incredibly short amount of time, this very capability made image classification using DL architectures one of the most studied technological applications of the COVID-19 pandemic. However, as the domain broadened, so did the degree of fragmentation. The reported levels of accuracy rose, at times reaching unreasonable values, while concerns regarding generalisation, bias in datasets, and clinician acceptability were accruing in the rear. The promise regarding COVID-19 diagnosis using DL was clear, but so were the contradictions.

It would be ideal if deep learning systems developed for diagnosis using medical images possessed three key characteristics: robust generalisation, the ability to distinguish between similar diseases, and interpretable reasoning processes that are tractable for a clinician. These systems would work regardless of imaging equipment, location, and disease prevalence and would integrate seamlessly into clinical settings [4]. This ideal state, however, has never been observed in practice. A significant number of COVID-19 image classification models developed early in the pandemic were trained on small, imbalanced, and heterogeneously collected datasets, sometimes scraped from publicly available sources [5]. Although transfer learning techniques have been used to address limited datasets, they have introduced

How to cite this article:

numerous problems in machine learning (ML), such as “stealthy” data leaks, overconcentration on dataset artefacts, and focus on non-pathology attributes. Many research works have tried to address these issues using ensemble methods, attention models, or CNN-long short-term memory (LSTM) combinations. Although such methods optimised metrics on benchmark datasets, they did little to address core problems associated with interpretability, reproducibility, and robustness across healthcare institutions. The immediate implication is an increased divergence between algorithm performance results and readiness within the healthcare industry [6]. The secondary implication relates to healthcare policies, expenditures, and overall acceptance of artificial intelligence (AI)-supported healthcare systems [7].

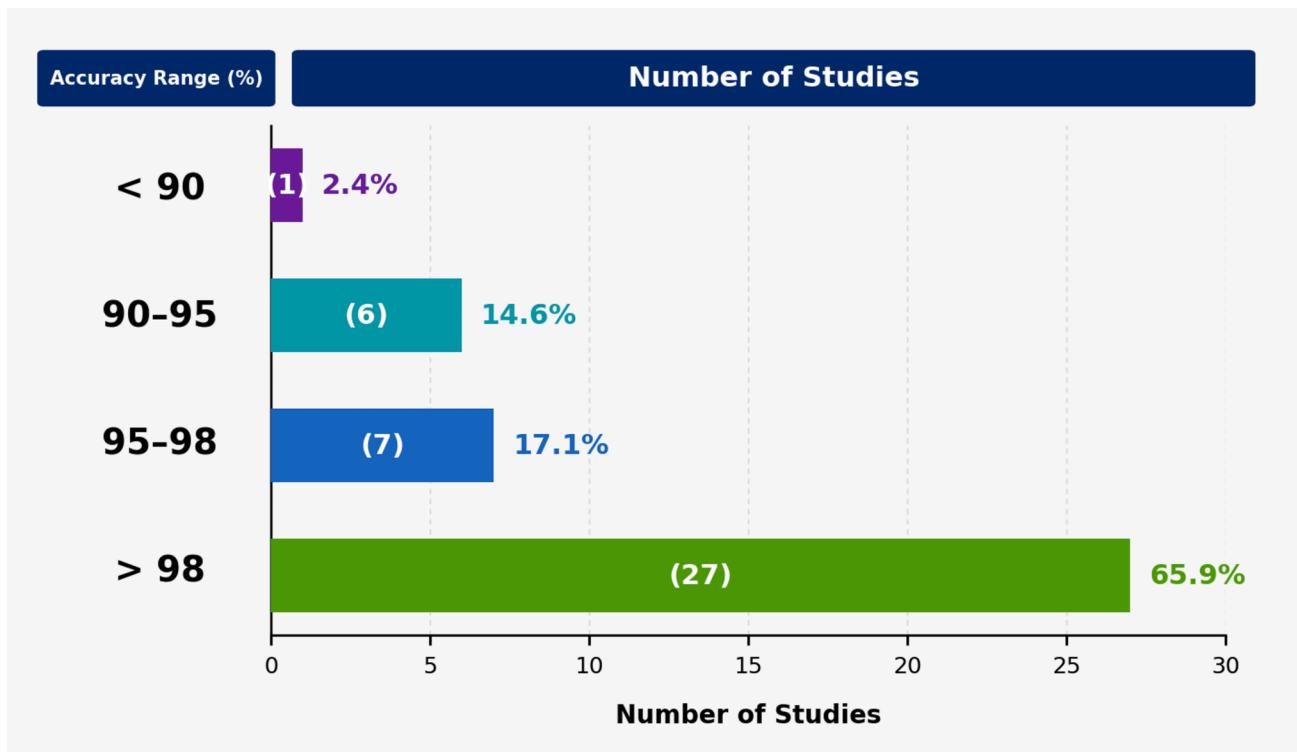


FIGURE 2: Distribution of Reported Accuracy Ranges

The graph in Figure 2 shows a clear concentration of accuracy results in the highest performance levels, with all studies claiming results above 98%. This clear concentration of results indicates a saturation of the metric in the literature, which raises questions about the over-optimisation of the metric on biased datasets. Although so much has been written, it is clear that little consolidated knowledge has been developed with respect to why this model works, how it is limited, and what the cost of this model is relative to alternative models. By examining what has been proposed in the literature as well as what has been overlooked and why this is important, this review aims to meet this need.

There are many methodologies that have been applied in COVID-19 image classification using DL [8]. There was an early attempt to use transfer learning on existing CNNs such as VGG, ResNet, DenseNet, and Inception to quickly apply them in the field with a small amount of training data. In subsequent studies, there have been attempts to use ensemble methods and segmentation-driven approaches to accurately locate the infected regions by minimising the background noise. Recent approaches involve the use of vision transformers and hybrid ConvNeXt-ViT architectures. While considerable progress has been achieved in deep learning methods for classifying COVID-19 images, a number of limitations exist in terms of methodology and translation. The previous literature reviews generally cover the description of CNN models and transfer learning strategies, as well as the performance measures of the models described in them, but do not properly assess the methodology and clinical significance of the developed models. Furthermore, there is also a lack of attention paid to other aspects such as dataset heterogeneity, data leakage, external validation, publication bias,

uncertainty assessment, explainable artificial intelligence (XAI), and reproducibility. Differentiating factors like imaging modality, dataset characteristics, preprocessing procedure, validation procedure, and evaluation metric preclude cross-comparisons between different studies. To overcome these drawbacks, this study proposes a systematic review of the use of deep learning techniques in the classification of COVID-19 medical images based on the PRISMA methodology, complemented by bibliometric analysis. Unlike previous reviews, which focus mainly on a summary of the architectures used and performance metrics, this review adopts a multi-dimensional assessment approach where not only the methodologies used but also their advantages and disadvantages, as well as potential applications to practice, are analysed. Apart from evaluating the deep learning architecture, imaging modality, dataset, and classification performance, this review also evaluates the quality of the dataset used, validation, external validation, XAI, reproducibility, publication bias, and translational readiness.

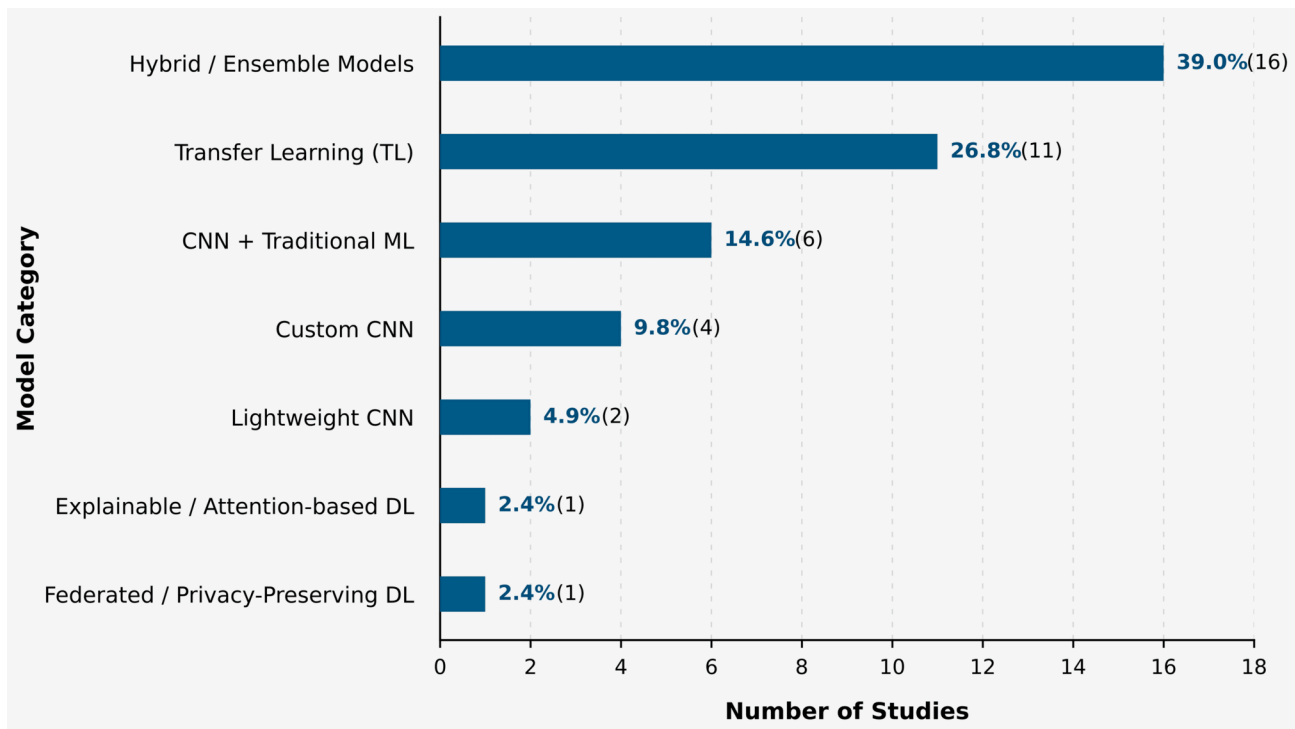


FIGURE 3: Model Architecture Categories in COVID-19 Image Classification

DL, Deep Learning; COVID-19, Coronavirus Disease; CNN, Convolutional Neural Network; ML, Machine Learning

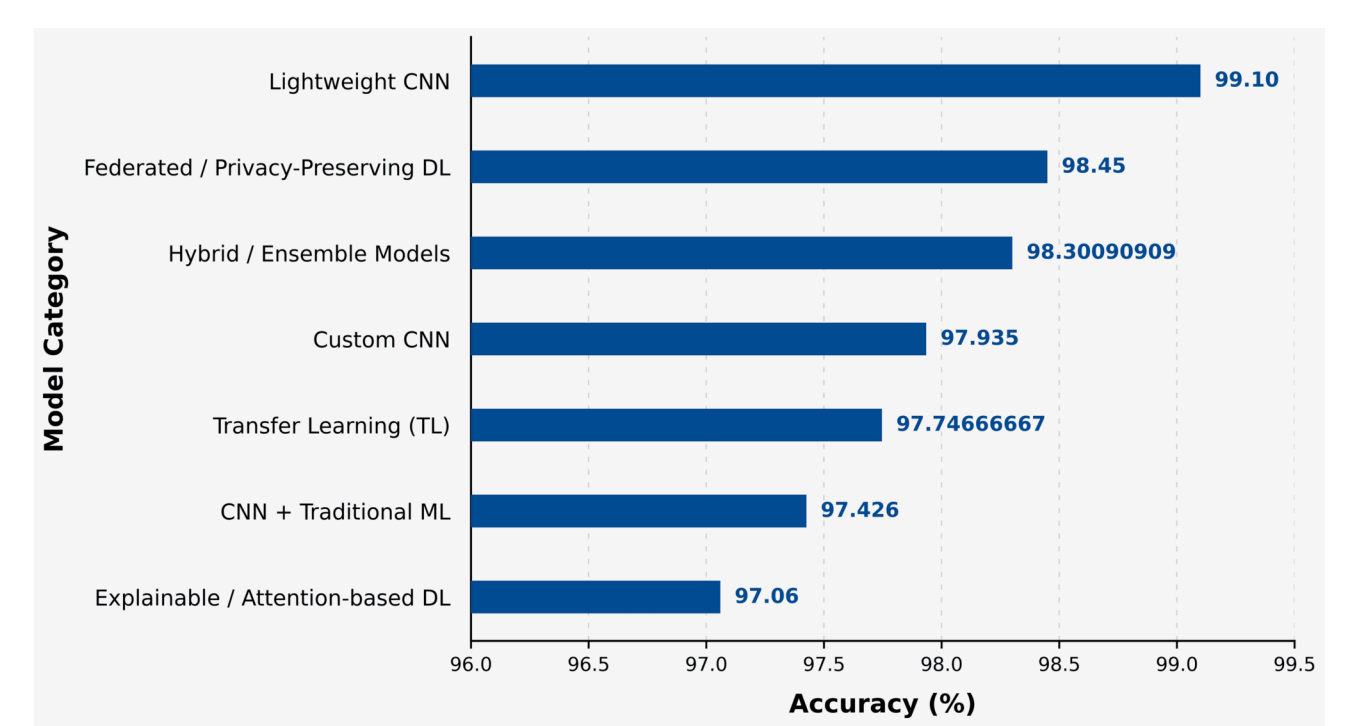


FIGURE 4: Average Classification Accuracy Across Deep Learning Model Categories

DL, Deep Learning; CNN, Convolutional Neural Network; ML, Machine Learning

Figure 3 reveals the prevalence of transfer learning and hybrids along with ensemble models in the literature as a means to achieve performance optimisation, whereas Figure 4 shows the emphasis on hybrid and ensemble-based approaches for attaining higher levels of classification accuracy. However, there are few studies that address the need for developing lightweight, explainable, and privacy-preserving models. Although the reviewed articles show substantial progress in technical advances, several research gaps can be identified in the area of medical image classification using deep learning concerning computational methods. First of all, many articles report performance comparisons using different metrics and different datasets and different preprocessing approaches. Thus, it becomes hard to understand which methodology has been the best-performing one in terms of accuracy. The use of Grad-CAM and attention mechanisms as explanation methods is common in the reviewed studies; however, they are not always used to systematically explain the rationale behind the model decisions. Federated learning and other methods of incorporating privacy preservation in medical image analysis using AI are poorly covered in the literature [9] and can be considered as areas for further research. Most of the reviewed articles lack systematic analysis of the discussed architecture performance in terms of reliability, validity, and applicability, as well as translation to practice. Besides Grad-CAM and attention mechanisms, there are other approaches to achieving explainability in the classification of COVID-19 chest X-ray and CT images reported in the literature. For instance, radiomics-based methods have been implemented using GLCM, GLRLM, GLSZM, and GLDM, combined with traditional ML classification algorithms, such as logistic regression, support vector machines, and LASSO. The literature also reports on the use of various explanation methods and tools, including LIME, Grad-CAM++, layer-wise relevance propagation, saliency maps, infected region contour visualisation, and feature importance heatmaps, among others. Although SHAP value explanation, occlusion sensitivity analysis, testing with concept activation vectors, counterfactual explanation, and many others are considered state-of-the-art in the realm of trustworthy AI, there are few studies on their use in the context of COVID-19 chest X-ray and CT image analysis. Based on the reviewed literature, it is possible to conclude that the implementation of explainable AI methods is conducive to obtaining transparent and trustworthy results in the analysis of medical images. Informed by the computational learning theory perspective, which is based on the fundamental principles of representation learning and generalisation, and the diagnostic decision-

How to cite this article:

making paradigm, this study explores the concepts of distribution shift, feature attribution, and clinical validity as applied to deep learning algorithms for the diagnostic analysis of medical images. The models under review are thus evaluated in terms of their capacity to offer reliable and consistent performance across different datasets, domains, and populations. From the socio-technical perspective, the review aims to position deep learning-based diagnostic analysis as a process influenced by both technical (network structure, depth, and parameterisation) and social (data acquisition, preparation, validation, and ethical review) factors [10]. In this regard, the current review tries to offer a comprehensive synopsis of the trends in deep learning research in medical image classification, which would serve as a basis for critical discussions on trustworthy AI in diagnostic imaging. Apart from identifying strengths, weaknesses, opportunities, and threats of applying deep learning to medical image classification in general and COVID-19 diagnostic analysis, specifically, this review attempts to move beyond simply comparing the accuracy performance of the algorithms under review. By focusing on articles that offer a comprehensive discussion of the technical and methodological aspects of the research, the current review tries to synthesise a critical appraisal of the deep learning literature for medical image classification in terms of validity and generalisation. For policymakers and healthcare administrators, this review aims to offer implications for regulating medical image diagnostic analysis using AI, as well as for promoting the development of reliable and trustworthy models for diagnostic imaging. From the medical professional's perspective, the review offers a critical analysis of how the majority of deep learning methods have been applied to medical image classification, thus helping choose the most appropriate approaches for diagnostic practice. For research practitioners, the current review will serve as a basis for developing new areas of trustworthy AI for medical image analysis with applications to pandemics. Based on the CARS model, the current research was initiated by defining the research territory, which has been achieved by outlining the role of deep learning in medical image classification during pandemics. As such, this article has identified the research niche by elaborating on the research gaps in the current literature and focusing on areas where validation, generalisation, and explainability of deep learning methods are critical issues, addressing the need for further research and development in this regard. The current systematic review thus responds to the needs of this niche by conducting extensive research on the development of trustworthy AI for medical image classification based on the critical analysis of the literature, as well as bibliometric analysis.

Review

In this study, the COVID-19 diagnosis using deep learning for image classification appeared not only as a technological innovation but also due to the pressing needs that became apparent in the course of the epidemic. Though molecular diagnostic procedures like RT-PCR were the gold standard in most clinical settings, their lack of scalability and the time required for the diagnosis to be realised made it imperative that alternative methods of diagnosis be made available [11]. Medical imaging, particularly the CXR and chest CT, emerged as a readily available alternative [12]. However, it relied heavily on the expertise of the interpreting clinician. Deep learning with the capacity to abstract hierarchical representations of images has been poised to unlock joint assessment of imaging studies [13].

How to cite this article:

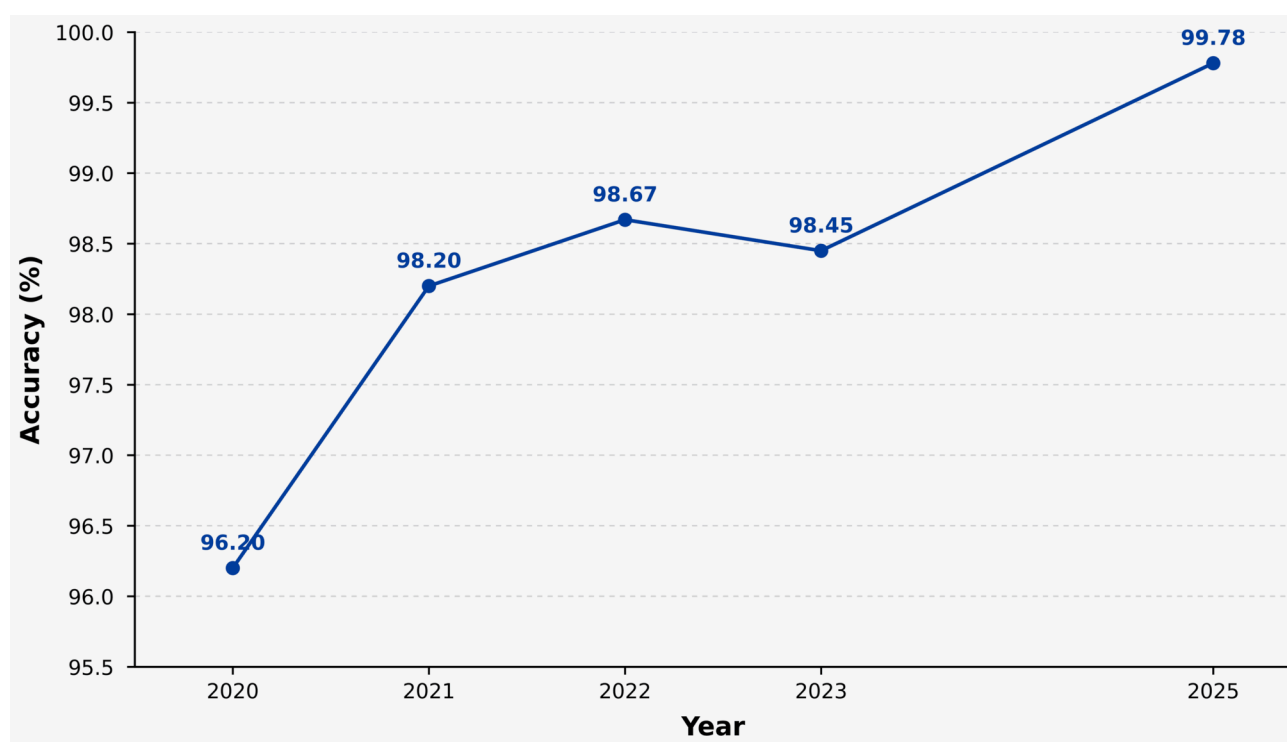


FIGURE 5: Year-wise Trend of Reported Accuracy in COVID-19 Image Classification

COVID-19, Coronavirus Disease

Figure 5 shows the trend of improvement in the average reported accuracy of deep learning-based COVID-19 image classification models over the years. Although the trend indicates progress in terms of architectural and data-related improvements, the saturation of accuracy values over the years also raises concerns about accuracy inflation. The importance of this domain is not only underscored by the level of known accuracy but also by how it relates generally to situations concerning emergencies, clinical support, and the potential applications of AI in healthcare systems. However, the rate of research advancement far exceeded that of methodological advancement, leaving a body of research literature with plenty of models but little rigour. A clear appreciation of what has been accomplished in the body of research and what has been left behind is crucial in determining the level of maturity of DL-based classification of COVID-19 into a useful diagnostic tool or an emerging research path [14]. In this research, the aim will be to critically assess the methods of using deep learning techniques for image categorisation to detect COVID-19 and specifically identify the methodological trends, inconsistencies, and research gaps based on the choices, methods, and findings regarding the limitations. Earlier relevant research mainly applied convolutional neural networks with techniques like transfer learning. Early studies used pretrained models such as ResNet, DenseNet, and VGG for COVID-19 classification [15-17]. The appeal of their approaches mainly arises from their efficiency and potential to attain a high degree of accuracy, often to the tune of over 95%, with minimal training images.

Although such works proved the existence of the possibility of diagnosis through DL, their pitfalls should not be underestimated. First, they are mostly limited to very small, carefully selected sets of data compiled from public sources, which are not always selected in a standardised fashion and are not collectively validated. Moreover, learning transferability from natural image sources such as the Internet-based ImageNet raises other queries in terms of applicability, as image textures derived from nonmedical images are unlikely to translate meaningfully to pulmonary pathology. Realising the medical requirement of separating COVID-19 from other respiratory infections, subsequent research work involved multiclassification models. Sethy and Behera [18] and Ibrahim [19] extended binary classification by incorporating pneumonia classes alongside normal cases, with promising classification performance. Due to

How to cite this article:

Ghosh S, Chhetri Y, Das K, et al. (September 21, 2026) Deep Learning-Based Image Classification for COVID-19 Disease: A Comprehensive Review. *Cureus J Comput Sci* 3 : es44389-026-00286-6. DOI <https://doi.org/10.7759/s44389-026-00286-6>

radiological similarities between COVID-19 and viral pneumonia, this research addressed an important medical concern. However, the analysis of the different models brings up a number of weaknesses. Most of the multiclass models' high accuracy was based on the fact that the classification was imbalanced, or the distribution of the labels was very simple. The issue of misclassification, as well as false positives, was not deeply evaluated by a few of the studies, especially when it comes to the differentiation between COVID-19 and other pneumonia. Therefore, though the multiclass model improved the realism of the diagnostic task, it did not have the necessary sophistication to provide a risk assessment. CT-based studies by Amyar and Zhou incorporated segmentation algorithms to identify infected lung regions before classification [20,21]. These models harnessed the increased resolution capabilities of CT scans to detect ground-glass opacities, and COVID-19 infection is indicated by bilateral involvement [22]. From a conceptual standpoint, segmentation classification pipelines are an advancement in that they integrate knowledge of anatomy in the learning process. However, the approaches using CT also bring in some new drawbacks. Compared with CXR, the availability of the CT scan is lower in many areas and also involves a higher radiation dose of exposure to the patients. Therefore, the use of the CT scan can be impractical in mass screening. Furthermore, there is an increased cost of pixel annotation in segmentation models due to the involvement of interobserver variability. Various hybrid architectures and ensemble learning concepts were also explored in different research works to improve performance and stability. Kanwal and Chandrasekaran used CNNs with bidirectional LSTM layers to capture spatial dependencies, while Ravi combined CNNs with meta-classifiers [22,23]. Performance improvements have been regularly reported with attention mechanisms and ensemble techniques, suggesting that architectural variety can buffer the shortcomings of individual models. These improvements frequently come at the expense of less interpretability and more complexity. Transparency problems are exacerbated by the fact that many of the ensemble models operate as black boxes piled on top of other black boxes. Furthermore, little research has assessed the statistical significance and clinical significance of performance gains. Despite the development of the hybrid approach, explainability and deployment were still unresolved problems for such research. External validation is one of the most critical problems in the literature. This aspect was highlighted by Aggarwal and Yang [24]. The reason is that models are trained on the data itself and not on disease-related patterns. Cross-validation across institutions is a very uncommon practice, while validating models across distribution shifts, like imaging modality, demographics, or disease prevalence, is even less common.

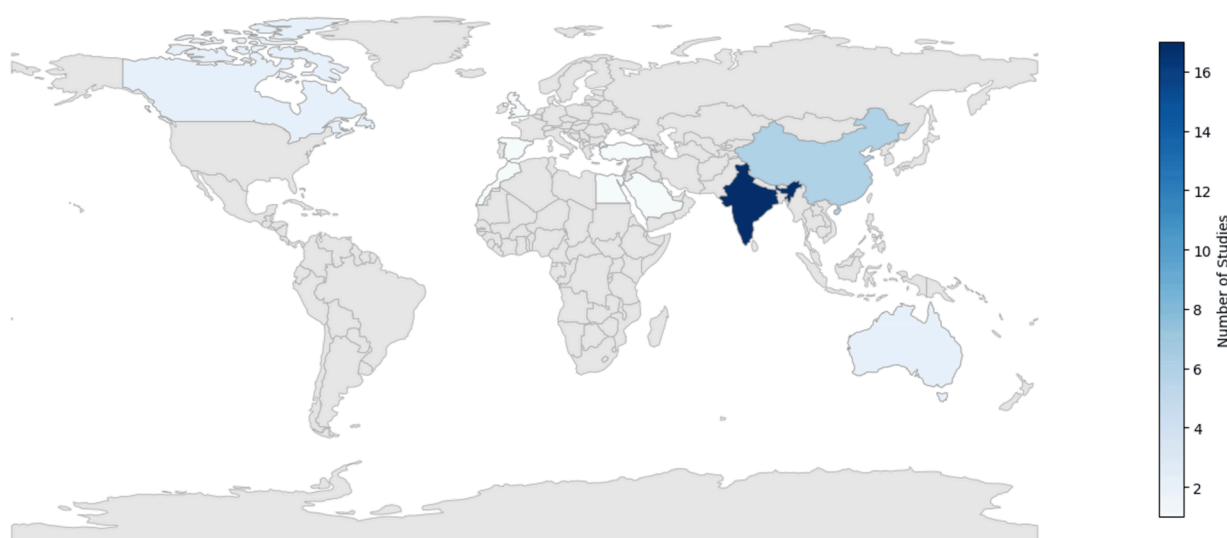


FIGURE 6: Regional Distribution of Research in COVID-19 Image Classification

COVID-19, Coronavirus Disease

How to cite this article:

Ghosh S, Chhetri Y, Das K, et al. (September 21, 2026) Deep Learning-Based Image Classification for COVID-19 Disease: A Comprehensive Review. *Cureus J Comput Sci* 3 : es44389-026-00286-6. DOI <https://doi.org/10.7759/s44389-026-00286-6>

This graph, Figure 6, represents the geographical distribution of research output in deep learning-based COVID-19 image classification studies. There is a concentration of research output in a few countries, which reflects the imbalance in global research participation. Although regions with better research infrastructure have more research maturity, there are regions that are not well represented. This reflects the risks associated with biased datasets, lack of external validation, and poor generalisability of the developed models. This drawback specifically counteracts the primary objective of a DL-based diagnostic method. This objective focuses on providing a dependable level of performance under practical diagnostic conditions. Lack of strict validation will render the reported accuracy values prone to misleading interpretations. Techniques like federated learning, focusing on maintaining patient privacy, were proposed as a method of overcoming data sharing limitations in healthcare. Malik demonstrated in 2023 that multi-disease classification using a federated learning framework is possible, providing directions towards successful collaborative learning without data aggregation. Techniques like these need much more literature research. This is a reflection of a larger divergence between the world of algorithm development and the reality of healthcare delivery systems. Throughout the literature, several themes have been identified. The preponderance of CNN transfer learning, ensemble/hybrid methods, and predominantly high accuracy (>90%). However, contradictions are also identified to be equally present throughout the literature, including high performance lacking high validation, a lack of interpretability despite increased architectural complexity, and a lack of clinical agreement despite increased variability within methods. Most importantly, the literature focuses on optimisation but neglects evaluation. Few studies investigate whether models are learning domain-relevant features, whether mistakes are being propagated properly in environments that mimic real-world scenarios, or whether uncertainty should be expressed to clinicians in certain forms. All these gaps indicate a lack of translational readiness despite technical achievement. On the whole, the literature portrays technical momentum but underdeveloped conceptual consolidation. Although other studies are in part pertinent to the objective of enhancing automated COVID-19 diagnosis, they are not adequate in dealing with the aspects of robustness, generalisability, and clinical consolidation. The lack of critical synthesis in the selected research makes it difficult to determine which developments are core and which developments are location-based. Additionally, this review fills this gap in research by examining deep learning methods using various perspectives. Hence, this study bridges the gap that exists in research by offering evidence-based guidelines on how deep learning methods in image classification should move from being in an emergency research phase to being in a stable research phase.

Methodology

The presented study employs a methodical bibliometric review approach that combines bibliometric scientific description and science mapping analysis with PRISMA-guided research selection. The goal of this approach is to provide a clear, organised, and repeatable synthesis of the existing scientific data about deep learning-based image classification for COVID-19 diagnosis. As stated in the present bibliography on bibliometric reviews, the current technique aims to uncover publishing typologies, research method patterns, and knowledge gaps throughout time, space, and model type, in contrast to a quantitative meta-analysis.

Review Design and Scope

The present study investigates deep learning-based image classification methods for COVID-19 diagnosis using a hybrid review strategy that combines bibliometric mapping and systematic literature review concepts. This review design is justified by the fact that COVID-19-related AI research is rapidly expanding and highly varied, making it difficult for even narrative synthesis to capture broad trends, methodological concentrations, and gaps. While systematic review criteria guarantee openness and repeatability of study inclusions, bibliometric analysis provides the mathematical foundation for tracking research activity. A structured search strategy was, therefore, implemented to identify relevant literature across a number of scientific databases with a focus on peer-reviewed studies that had been published during the period of the COVID-19 pandemic. Iterative refinements in the search terms allowed the retrieval of studies explicitly focused on deep learning techniques applied to chest X-ray and/or CT images within the context of COVID-19 detection and classification. Regarding the metadata, extracted information included publication year, country of origin, modality of imaging, and methodological focus to further assist in bibliometric and thematic analysis. In order to ensure methodological quality and traceability, study screening and selection were carried out in accordance with PRISMA guidelines. The titles and

How to cite this article:

abstracts were filtered using predetermined inclusion and exclusion criteria after duplicate references were eliminated. The possibly suitable papers underwent final full-text screening based on reporting quality and methodological relevance. A PRISMA flow diagram in Figure 7 depicts the entire article selection process, including the steps of identification, screening, eligibility, and final selection. The final list of studies for the inclusion phase was then described using bibliometric mapping. The distribution of studies geographically within the study domain was depicted after author connections were taken into account while allocating the country-wise publishing of the studies. Additional maps were created to show the trend of publications over time and research classified according to picture type, task complexity, validation type, and methodological focus. In order to analyse the coverage of studies throughout the research region without introducing any bias during the methodological stage, the maps were ultimately presented using descriptive tables and figures.

Data Sources and Search Strategy

To ensure methodological transparency, reproducibility, and comprehensive coverage of the extant literature, this systematic review was conducted in strict adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. The five world-renowned scientific databases, which include the Web of Science Core Collection, Scopus, PubMed, IEEE Xplore Digital Library, and ScienceDirect, were subjected to a systematic search. These platforms were chosen for their extensive indexing of multidisciplinary research in fields linked to deep learning applications in COVID-19 image processing, including computer vision, artificial intelligence, biomedical imaging, and medical informatics.

The search period, which ran from January 2020 to December 2025, coincided with the rapid development of deep learning methods for COVID-19 diagnosis. To preserve methodological consistency and scientific rigour, only English-language peer-reviewed journal publications were included. In order to guarantee the inclusion of excellent, empirically supported research, exclusion criteria included editorials, commentaries, book chapters, narrative reviews, theses, patents, conference proceedings, abstracts, preprints, and non-English publications.

Using the logical operators AND & OR, a strict Boolean search framework was developed by combining natural language keywords with restricted terms. COVID-19, deep learning, medical imaging, and diagnostic classification were the four key topic domains around which the plan was organised. The general search term used in all databases was:

("COVID-19" OR "SARS-CoV-2" OR "Coronavirus") AND ("Deep Learning" OR "Convolutional Neural Network" OR CNN OR "Transfer Learning" OR "Artificial Intelligence") AND ("Chest X-ray" OR CXR OR "Computed Tomography" OR CT OR "Medical Imaging") AND ("Classification" OR "Diagnosis" OR "Detection")

Despite keeping the core principles of the search rationale the same, the search syntax was adjusted for each database to fit its indexing criteria and query requirements. Backward citation searching was conducted for review articles and primary studies that received many citations in an effort to identify relevant studies that might have been overlooked during automated searches. Before the initial screening of titles and abstracts, any duplicate entries due to cross-database overlaps were identified and eliminated. The PRISMA flow diagram (Figure 7) shows the entire search approach, which includes retrieval statistics for each database, eligibility assessments, screening procedures, and selection of included items.

Literature Retrieval Process

PRISMA 2020 guidelines were strictly adhered to in the process of literature searching and study selection, which consisted of four steps: identification, screening, eligibility, and inclusion. Based on the pre-established Boolean search approach, bibliographical references were systematically acquired in the first step from the Web of Science Core Collection, Scopus, PubMed, IEEE Xplore Digital Library, and ScienceDirect. Duplicate references as a result of the overlap in databases were removed before moving on to the screening step. Subsequently, several authors separately analysed the titles and abstracts of selected articles using the predetermined inclusion and exclusion criteria. All papers that did

How to cite this article:

not address the issues associated with COVID-19 diagnosis, medical imaging, and image classification through deep learning techniques were excluded. Finally, full articles of those studies that could be considered relevant for the systematic review were assessed.

An article was considered eligible for selection based on the criteria that it should (i) either present or evaluate deep learning approaches for the diagnosis of COVID-19 from chest X-ray and/or computed tomography (CT) images; (ii) have a detailed methodology description and quantifiable performance measures; (iii) employ either publicly available or clinical imaging datasets; and (iv) be a peer-reviewed journal article in the English language between January 2020 and December 2025. Exclusion criteria encompassed duplicate reports, reviews, editorials, commentaries, book chapters, theses, patents, conference papers and abstracts, non-peer-reviewed preprints, works without sufficient empirical validation, and studies not centred on image-based diagnostic approaches for COVID-19.

The final corpus of included studies underwent bibliometric analysis, critical appraisal of methodological quality, assessment of dataset characteristics, comparative evaluation of model performance, and thematic synthesis. Discrepancies in eligibility judgements were resolved through iterative discussion and consensus among the research team following independent reevaluation, thereby mitigating risks of selection bias. The PRISMA flow diagram (Figure 7) shows the whole study selection process, including identification, screening, eligibility determination, and final inclusion.

The literature retrieval and study selection procedures used in the systematic review are described in Table 1. A total of 1,138 records were found in the first search; 775 of them were screened, and 701 of them satisfied the first eligibility requirements for further analysis. In all, 46 studies were finally chosen to be included in the final systematic analysis after the planned inclusion and exclusion criteria were applied. The table provides a clear overview of the transition from the original corpus of identified literature to the final set of included studies by methodically outlining each step of the selection process.

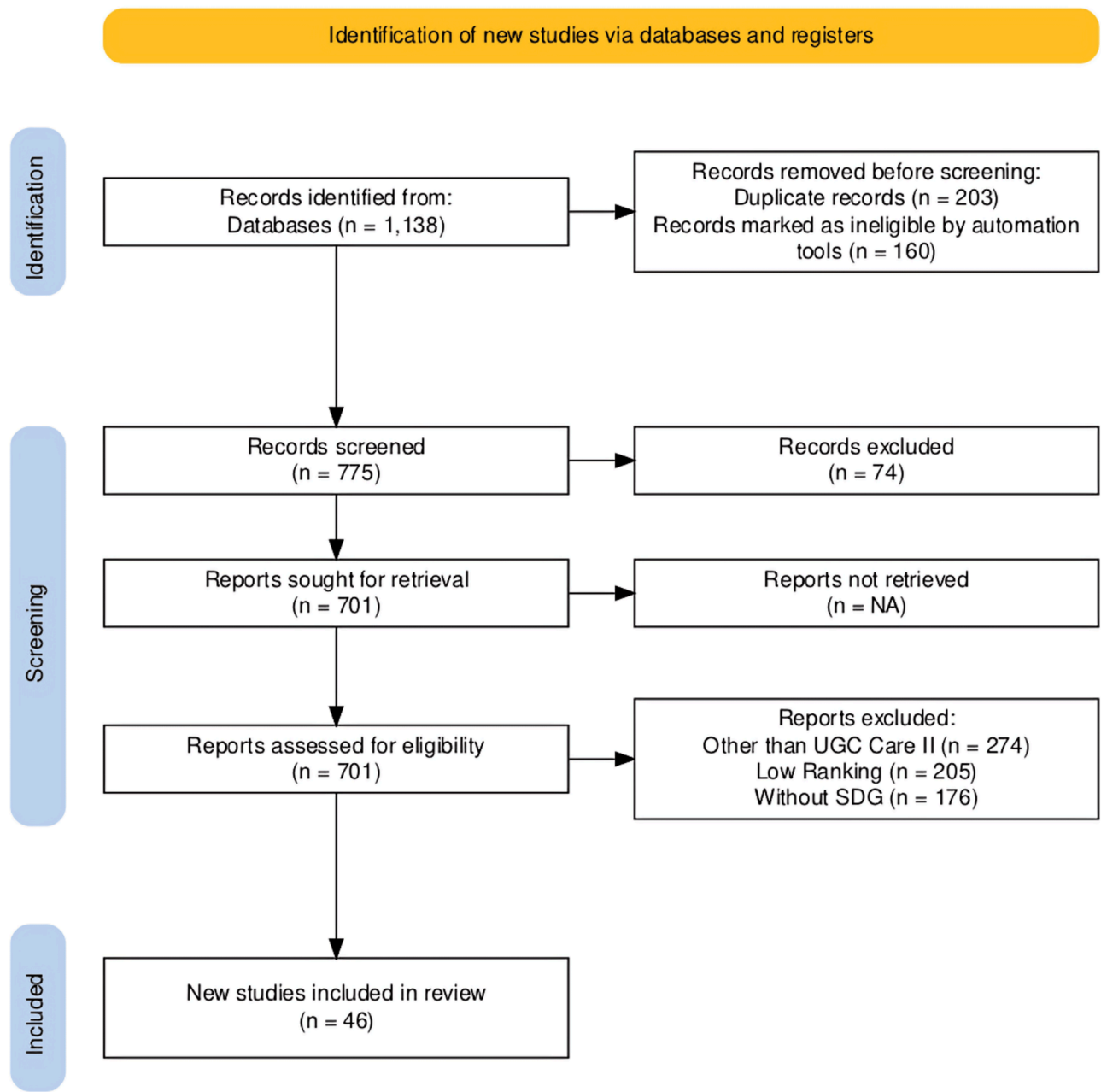


FIGURE 7: PRISMA Flow Diagram of Study Selection

PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; UGC, University Grants Commission; SDG, Sustainable Development Goal

How to cite this article:

	Database	Search Period	Initial Records Retrieved	Duplicate Records Removed	Records Screened	Full-text Articles Assessed	Studies Included	
	Web of Science Core Collection	2020–2025	250	70	180	160	10	
	Scopus	2020–2025	290	90	200	180	12	
	PubMed	2020–2025	180	55	125	115	8	
	IEEE Xplore Digital Library	2020–2025	150	48	102	95	7	
	ScienceDirect	2020–2025	268	100	168	151	9	
	Total	-	1,138	363	775	701	46	

TABLE 1: Database-wise Literature Retrieval Statistics

In Figure 7, a PRISMA flowchart illustrates the whole identification, screening, eligibility evaluation, and inclusion process. By indicating the number of records included and eliminated, the flow chart improves the study's methodology and provides a clear understanding of the whole selection process.

Inclusion and Exclusion Criteria

Before the full texts were analysed, inclusion and exclusion criteria were determined to guarantee consistency in techniques and analysis of relevance. These criteria were designed to reject studies that go beyond the conceptual bounds of this review and to include works that substantially advance our understanding of the idea of image classification in COVID-19 diagnoses using deep learning techniques.

Inclusion criteria:

A study was deemed eligible for review inclusion if it fulfilled the following requirements:

1. The primary research design strategy used in the study was deep learning techniques, such as convolutional neural networks or their variations.
2. The study focused on image-based diagnosis and classification with regard to COVID-19 diagnosis, specifically using CT and X-ray images of the chest.

How to cite this article:

3. Because the work was published in a peer-reviewed journal, the methodology and results were assured to be of a high calibre.
4. There was enough methodological material to allow for a critical analysis, including details regarding models, datasets, and measurement criteria.
5. The article was released in English during the designated period of the epidemic.

Exclusion criteria:

Articles that had any of the following elements were eliminated from consideration:

1. Neither medical imaging data nor non-imaging modalities of diagnosis, such as clinical, genomic, or epidemiological modelling, were the focus of the study.
2. Deep learning was not used in this work; instead, conventional ML techniques were employed.
3. Without complete experimental verification, the publication is a type of conference paper, preprint, editorial, review, or brief communication.
4. Reproduction and comparison were not made possible by the lack of methodological transparency.
5. The suggested work was beyond the scope of research for COVID-19 image categorisation, such as studies on various respiratory illnesses that were not particularly concerned with COVID-19 evaluation.

By using these criteria, it was ensured that the screening and assessment procedures were consistent, and in the end, the research included in this review satisfied the methodological and focus requirements.

Bibliometric Analysis Procedure

A bibliometric study was conducted to methodically identify and characterise the research environment encompassing COVID-19 image categorisation by deep learning following the conclusion of the final selection procedure for qualifying studies. In order to construct an organised tabulation for simple descriptive analysis and visualisation, the bibliographic information was extracted from the published research publications.

In order to assess the dynamic trend of publications during the pandemic, the publication year variable was gathered. This variable might be used to identify the trend and publishing bursts. The authors' institutional affiliations were used to determine the regional distribution. Based on the associated author's initial affiliation, each study was allocated to its nation. The criterion was applied in order to preserve equality between the various nations.

Articles have been organised according to their relation to certain characteristics of the suggested methodology in addition to the time and place markers. These characteristics included the kind of classification task (binary vs. multi-class classification problems), validation techniques, image modality (chest X-ray vs. computed tomography imaging), and whether explainability components were taken into account.

Conventional bibliographic processing and visualisation tools were used to enable bibliometric mapping and descriptive analysis, producing figures and tables that showed research coverage, publishing trends, and regional distribution. The selection of studies and the evaluation of their quality were unaffected by the bibliometric outputs, which were only utilised for descriptive and analytical purposes. This prevented methodological bias from being introduced and guaranteed that the bibliometric process served to objectively characterise the literature.

Thematic Classification of Studies

The included studies were divided into topical groups based on their main methodological focus to facilitate organised comparison and synthesis. The research objectives, model structures, and experimental designs were repeatedly examined to derive themes using an inductive categorisation technique. Every study was categorised into one or more topic groups based on its primary contribution.

The thematic framework consisted of the following categories:

How to cite this article:

1. CNN-based classifiers.
2. Transfer learning approaches using pretrained architectures.
3. Hybrid and ensemble models combining multiple deep learning components.
4. Frameworks that focus on explainability using attention models or saliency maps.

Those papers that were devoted to other issues such as dataset creation, data preprocessing, or the creation of the experimental protocol were classified according to the main focus of the analysis. The accuracy of these classifications was verified according to their consistency both with the methodology used in the paper and with its objectives. In case one paper discussed several methodological aspects, the classification was made depending on the most important methodological aspect for reaching the experimental objectives. This thematic classification could be useful for carrying out a comparative analysis of the strengths and weaknesses of various deep learning approaches to image classification of COVID-19 (Figure 8).

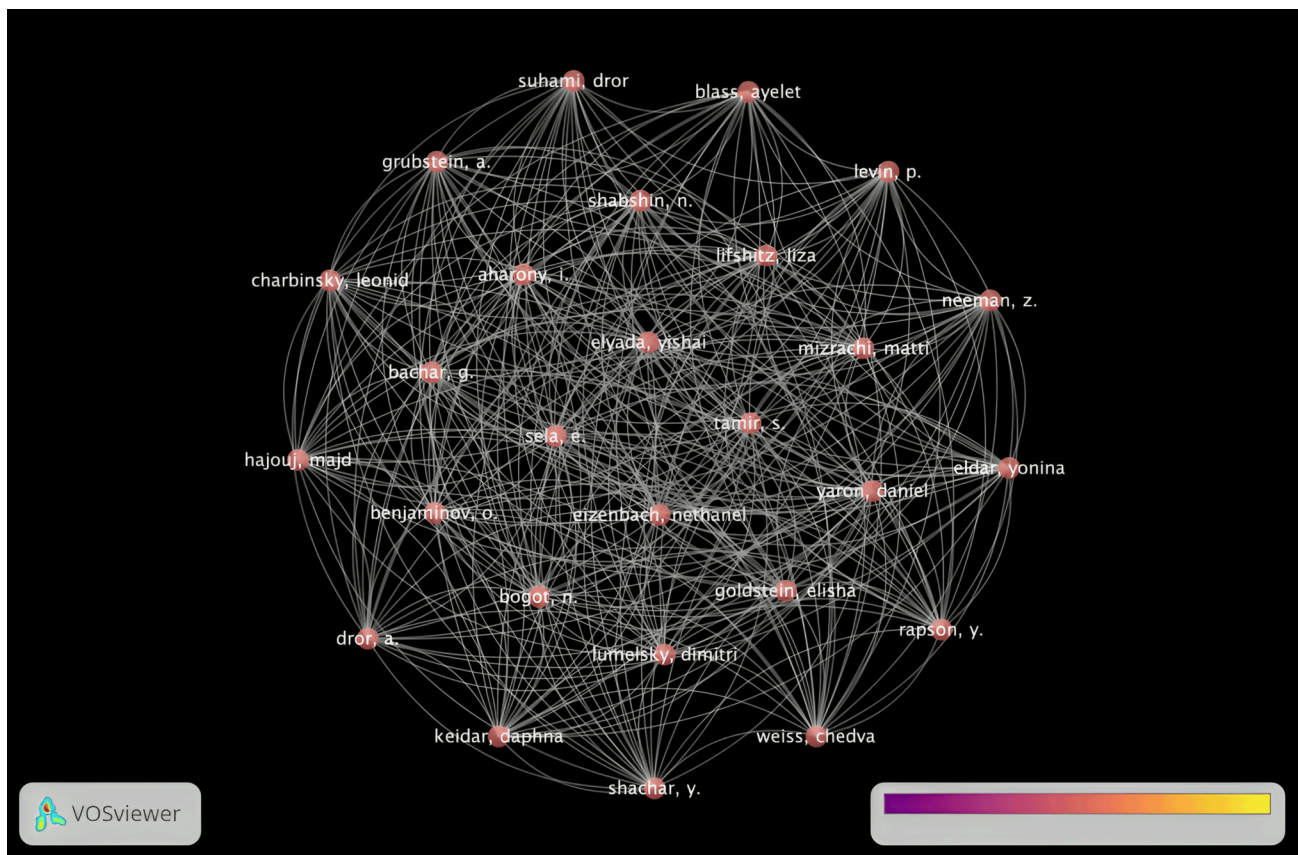


FIGURE 8: Author Co-authorship Network Visualized Using VOSviewer

The density of connectivity reveals that there is a domination by a few interlinked research clusters.

Quality assessment and bias consideration:

A systematic evaluation of the chosen papers was conducted even though this research endeavour is not meant to be a quantitative meta-analysis study. This was done in order to offer a critical assessment of such research in terms of methodological soundness and potential bias sources. This activity's goal is to give a contextual understanding of the chosen outcomes.

How to cite this article:

Both studies have been evaluated on a number of quality-related factors, including the adequacy of the datasets, design, validation, and reporting transparency. Since these factors are crucial in establishing the dependability and generalisability of deep learning models, the dataset sufficiency, class distribution, and source have all been taken into account. Small datasets with large augmentations, especially those with extraordinarily high accuracy values without validation, have been documented.

Institutional homogeneity, validation techniques, and dataset overlap were all potential causes of bias. Due to possible techniques for label prediction employing overlapping datasets, studies utilising single-source datasets or non-independent train-test splits were identified as biased. Furthermore, since the models could be detecting statistical patterns unique to their datasets rather than actual patterns, the absence of validation on datasets other than their own was also regarded as bias.

The approach for evaluating the research's quality also addressed the problem of model interpretability. Studies that include explainability techniques, like attention maps, or explanations, such as saliency visualisations, were emphasised due to their potential to increase clinicians' trust. Despite their accuracy, the performance-only experiments that ignored model interpretability were viewed as having little promise for translation.

This review determined the level of analysis in terms of weightage, in contrast to many studies that utilise exclusion criteria linked to quality levels. Research that has verified its techniques and been highly transparent was given more weight, whereas results from lesser studies were examined cautiously. This has enabled critical examination to be carried out while upholding inclusion.

Results

Methods and Model Architectures

In order to handle small COVID-19 datasets, the majority of the research has employed CNN-based architectures, including DenseNet, ResNet, VGG, Inception, and Xception, frequently utilising transfer learning [16]. Classification performance was further enhanced by hybrid models that combined CNNs with LSTM, attention mechanisms, or ensemble learning [23].

Metaclassifiers and federated learning have not been extensively studied to improve privacy and generalisability.

Performance-based metrics: Regarding the categorisation of COVID-19 from chest X-ray and CT images, accuracy rates recorded in various research often exceeded 95%, with some studies reporting an accuracy of 99% or greater in binary classification issues [17]. Similar to the binary classification results, the F1 and AUC scores for multiclass classifications (i.e., COVID-19, pneumonia, and normal) were more than 0.9.

Data source and pre-processing: Datasets, which had anywhere from several hundred to thousands of photos, were usually collected from open sources such as GitHub, Kaggle, and hospital databases [18,21]. To mitigate issues related to insufficient data and make the model more resilient, data processing methods and GANs were employed.

Comparative Performance Analysis

Summary of Findings: According to the reviewed research, deep learning-based diagnostic systems perform at a very high level, as there are cases when COVID-19 was diagnosed on chest X-ray or CT scan images with over 95% accuracy. However, as total accuracy is highly dependent on the frequency of positive instances, data quality, image type, and other variables, a direct comparison of performance is not always simple. Consequently, accuracy alone is insufficient to assess diagnostic performance, even if it is a sign of the model's capacity to function properly.

Nevertheless, the majority of research shows that overall accuracy is not as useful as sensitivity and specificity. While the latter lowers false positives and guarantees that healthy people are not exposed to further testing and the related costs and discomfort, the former measure must be high to limit false negatives and guarantee that sick patients are not overlooked. Excellent findings are frequently obtained on specialised datasets rather than populations that are generally relevant, despite the fact that numerous articles indicate both sensitivity and specificity above 95%. The area under the

How to cite this article:

ROC curve (AUC), a useful statistic that assesses a model's capacity to accurately differentiate between healthy and sick individuals, is also used in a number of the examined research publications to assess performance. However, the use of comparable public datasets rather than large multinational populations may be partially responsible for the consistently high AUC values obtained by several research studies. In addition, few articles report confidence intervals for AUC, which complicates the comparison between algorithms.

Similarly, precision, recall, and F1-score allow researchers to assess model performance from slightly different perspectives, which is particularly useful given the potential variance between populations of healthy and sick patients. As a result, comparable performance across various F1 scores is even more informative than overall accuracy, especially for positively skewed populations. The reviewed articles demonstrate that multiple investigators have been able to achieve consistent and high performance across a variety of scans, imaging types, and populations. However, only a few studies compared their solutions against each other and evaluated diagnostic consistency in terms of F1 scores.

Remarkably, none of the investigations explicitly mention calibration, which limits the understanding of how informative the probability scores from the models are in practical applications. In other words, well-calibrated models produce results that are intuitive for clinicians, which facilitates the incorporation of machine learning findings into medical practice. Lack of calibration, in turn, limits model utility by producing confusing evidence, essentially rendering it useless for practical implementation. Thus, the point here is that the fact that a relatively small number of the reviewed studies discuss calibration is reflective of the insufficient attention given to this essential element of modelling.

Likewise, a relatively limited number of papers have tried to estimate the uncertainty in the diagnostic predictions. The ability to identify uncertain cases becomes more and more relevant, especially considering the field of medical imaging, which requires robust automation. In general, the vast majority of scientific papers do not provide an explicit definition of whether the estimated label can be considered certain. In recent years, a great amount of effort has been made in order to implement the capability of estimating uncertainty into deep neural networks using MC dropout, Bayesian models, deep ensembles, and other means. It would be useful if future research considered this approach.

To summarise, the evaluation of generalisability is hindered by the problem that in most cases, the articles under analysis focus solely on internal validation. Several papers note that their models have been trained on a few large datasets but do not give examples of how the results were validated at other facilities. As a result, while diagnostic accuracy is frequently reported, it is difficult to estimate whether such accuracy is consistent across different populations or if it benefits from access to the same data used for training. Therefore, it is recommended that future studies explicitly report validation strategies employed to confirm the results.

Conclusion

The reviewed studies overwhelmingly emphasise accuracy, sensitivity, and specificity as the key parameters for evaluating deep learning-based diagnostic models for medical imaging. Several investigations demonstrate that these metrics can routinely exceed 95% for COVID-19 detection. However, the current literature has limited comparisons between models, which should be expanded in future studies. In addition, while most manuscripts frequently report excellent performance across a variety of data, populations, and imaging types, it is unclear how their results generalise to different institutions. Moreover, only a small number of investigations explicitly mention calibration and uncertainty estimation, which are both crucial for implementing such models in medical practice, particularly in the early stages of development. Therefore, future research should incorporate additional metrics and explicitly report calibration and uncertainty estimation techniques, if employed.

Limitations and Challenges

Some of the major hurdles that have been pointed out include the lack of balanced and limited datasets, overfitting, lack of external validation, and issues related to interpretability. Generalisation to different groups and machines has been discussed, along with the incorporation of AI models in clinical settings [24-26].

How to cite this article:

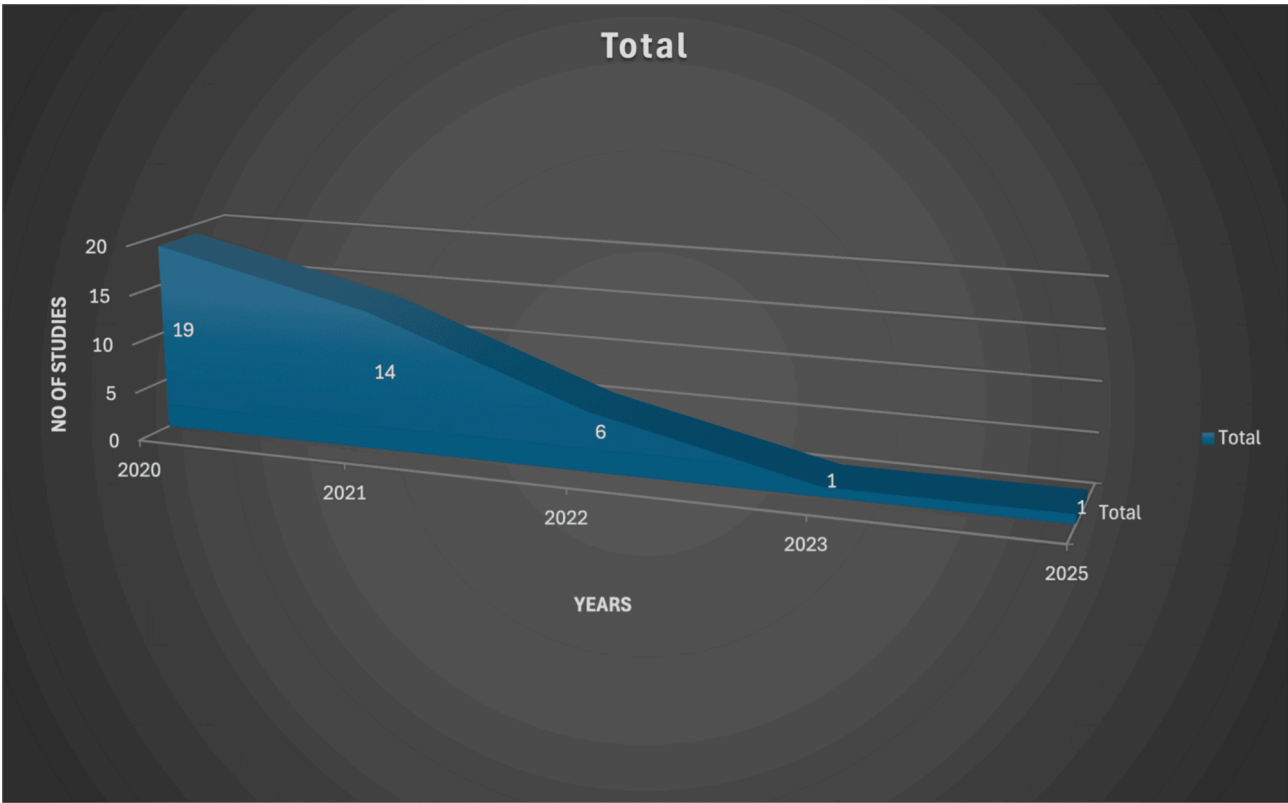


FIGURE 9: Temporal Trend of Influential COVID-19 Image Classification Studies (2020–2025)
COVID-19, Coronavirus Disease

Figure 9 illustrates the temporal dynamics of scholarly output in deep learning-driven image classification for COVID-19 from 2020 to 2025. A notable increase in research activity in the initial stages of the outbreak is observed in the publishing pattern, followed by a considerable change in the focus of the research subjects. Considering that most of the research is centred on convolutional neural network architectures, approaches to transfer learning, and publicly available chest X-ray and computed tomography databases, the drastic growth in the number of publications in the periods of 2020 and 2021 can be explained by the urgency of finding an automated solution to the problem at hand. In this phase, attempts at showcasing the technical possibilities of deep learning models in the context of image-based diagnostics were a key characteristic. Starting from 2022, the field of study has already demonstrated the diversity that goes beyond optimisation and architectural designs of the model itself. An increasing number of papers deal with such issues as the robustness of the model, its explainability, its ability to generalise across different populations, external validation, uncertainty estimation, and implementation within the clinical workflow. The topics of explainable AI, multimodal data fusion, federated learning approaches, and privacy-preserving methods are becoming increasingly common, which indicates the increasing attention to research topics that prioritise issues of a systemic, operational, and ethical nature over the issue of classification accuracy. Therefore, the diagram in Figure 9 should be viewed as an indicator of an emerging research trajectory, rather than a measure of the quantitative growth of the literature. The chronological distribution of the research demonstrates a shift from the hasty proof-of-concept experiments which were initially carried out during the outbreak to more rigorous and replicable science with practical applications. From enhancing the performance of predictions to solving the problems related to practical and epistemological considerations when using reliable and applicable systems of deep learning for diagnosing patients with COVID-19, this development is

characterised by a broadening of academic interests. In any case, Figure 9 demonstrates the main message of this research paper, which is that progress in robustness, explainability, generalisability, validation, and clinical feasibility should be taken into consideration along with improved performance in medical image analysis for COVID-19.

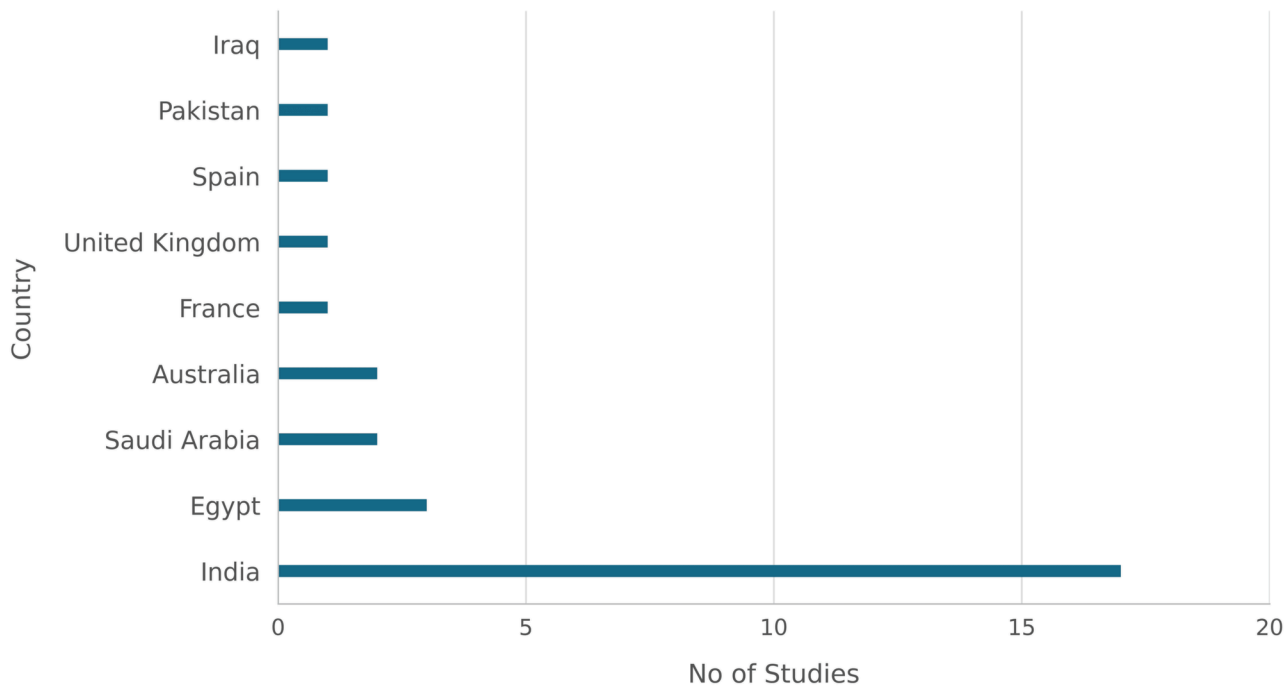


FIGURE 10: COVID-19 Imaging Classification Publications Based on Countries (2020 to 2025)

COVID-19, Coronavirus Disease

Figure 10 shows the geographical distribution of the scientific articles on COVID-19 image classification through deep learning from 2020 to 2025. From the analysis of the data, there is a geographical concentration of research production in a very few countries, where India contributes the most (17 articles), followed by China (6). Australia, Canada, Saudi Arabia, and Turkey have contributed two papers each, while Egypt and Iran have three each. Bangladesh, France, Iraq, Israel, Morocco, Pakistan, South Korea, Spain, and the United Kingdom have each contributed one article. The geographical distribution above demonstrates the stark disparity in research production quantitatively. China has contributed about 13.0% of the total 46 research papers, while India has contributed more than 37.0%. Taken collectively, the two countries contribute half the body of research being considered here, indicating that a large proportion of recently produced research emanates from a narrow geographical location. It shows an extremely uneven geographical distribution of intellectual labour despite the contribution of several other countries to research. The external validity of the deep learning models could be at risk due to the regional differences in the patient population, medical facilities, imaging modality, epidemiology of the disease, characteristics of the dataset, and the clinical process involved. Therefore, the aforementioned distribution indicates the requirement for multi-site data collection, geographic diversity in sample selection, and rigorous external validation in different clinical scenarios in order to ensure generalisability and applicability of COVID-19 diagnosis tools using deep learning technology. Overall, Figure 10 demonstrates that besides academic considerations, there is much more about the uneven distribution of the research that needs to be considered.

Discussion

The purpose of this research was to critically analyse the current status of deep learning-based image classification for COVID-19 diagnosis, taking into consideration methods of data usage, thorough validation processes, improvements in architecture design, and translational readiness. From the results of this study, one can see that there are many

How to cite this article:

innovations and improvements, but also many challenges that remain consistently present in the practice. Based on the theoretical framework used in this critical analysis, the implications of this research will be discussed within the context of the related literature.

Interpretation of Core Findings in Relation to Prior Research

This review aimed to evaluate the efficiency of deep learning approaches in image classification for COVID-19 diagnosis. Model architecture choice, type of imaging data used, data processing methods, validation approaches, and clinical translation challenges were among the central themes covered. The review findings demonstrate that the field has made significant progress in terms of applying deep learning to classify medical images for COVID-19 diagnosis [27]. In particular, the use of CNN-based approaches, including transfer learning, has been widely explored and has demonstrated classification accuracy of over 95% for both CXR and CT images. This is largely consistent with prior studies that have demonstrated the potential of using pretrained deep learning architectures for automatic diagnosis of COVID-19 [28-30]. Furthermore, the high level of performance demonstrated by multiple studies suggests that deep learning approaches can identify patterns in medical images associated with COVID-19 [31-34].

Architectural Trends and Theoretical Implications

The predominance of CNN-based architectures, including ResNet, DenseNet, and VGG families, can be explained by their effectiveness in learning localised visual representations from medical images [35-37]. Convolutional operations are particularly suitable for identifying spatial patterns associated with pulmonary abnormalities, including ground-glass opacities and consolidation [38-40]. The findings of this review therefore support the importance of representation learning in automated medical image analysis and help explain the continued adoption of CNN-based approaches [41-43]. By enabling the adaptation of pretrained architectures to medical imaging problems with relatively little training data, transfer learning has significantly accelerated the development of COVID-19 classification models [44-46]. However, the features of the source and destination domains determine how beneficial transfer learning is. Clinically significant pulmonary characteristics may not always be clearly correlated with features obtained from natural-image datasets [47-49]. As a result, proper dataset development, domain-specific validation, and meticulous evaluation of learnt representations should not be replaced by transfer learning. The report also noted that hybrid and ensemble techniques are being used more often. The instability of individual classifiers may be lessened, and robustness may be enhanced by combining many models or complementing learning components [50-55]. However, improved clinical effectiveness is not always correlated with increased architectural complexity [56-58]. Increased computing demands and decreased interpretability might result from more intricate structures. Because they may model global contextual information and larger spatial linkages, Vision Transformer and attention-based techniques offer an emerging trend [59,60]. However, the studies under discussion do not demonstrate that such approaches outperform CNN-based architectures in the case of very limited training data sets. Such a finding emphasises the importance of evaluating architectural complexity in terms of clinical significance, interpretability, and validation, along with dataset appropriateness.

Comparative Performance and Clinical Interpretation

While an accuracy of 95% was frequently reported in the studies, accuracy alone is insufficient for assessing the clinical reliability of diagnostic models. Disease prevalence, class balance, and the nature of the dataset itself can greatly influence the level of accuracy, especially in the case of unbalanced classes. Hence, a more in-depth analysis should consider the clinical application of the model and its validation alongside sensitivity, specificity, AUC, precision, recall, and F1-score. Since false-negative predictions can lead to delays in the diagnosis and treatment process, as well as further propagation of the disease, sensitivity is particularly important in screening applications. Conversely, false-positive predictions can lead to unnecessary isolation, additional laboratory tests, increased costs, and psychological stress. Due to the unequal impact of those types of errors in a clinical setting, the application of the model should define the proper ratio between sensitivity and specificity. Discrimination at different classification levels is another perspective provided by AUC. Still, high AUC values obtained from internal validation or homogeneous datasets might overestimate the true diagnostic performance of the model. In a similar vein, when class distributions are unequal, accuracy, recall, and F1-score can offer more insightful evaluations. Therefore, rather than depending just on accuracy, future research should present a

How to cite this article:

standardised set of clinically significant performance metrics, according to the comparative evidence. Uncertainty in prediction and calibration is another crucial factor. A model may produce probability estimates that are inaccurate in reflecting the true chance of illness, even when it achieves good discrimination. Therefore, inaccurate levels of confidence in clinical decision assistance might result from poorly calibrated forecasts. In a similar vein, uncertainty estimates can assist in differentiating between situations that need more expert evaluation and those for which the model has substantial evidence. A significant gap exists between algorithmic performance evaluation and reliable clinical deployment due to the examined literature's scant attention to these factors.

Dataset Quality, Validation, and Generalisability

Moreover, it is evident from the results that the process of validation, as well as the characteristics of the data used, play an important role when considering the credibility of the results obtained. The performance of the model can be influenced by the use of small samples, imbalanced classes, multiple acquisition procedures, and different kinds of preprocessing. In addition, the utilisation of public repositories results in repetitive use of similar data in future research. It is especially crucial to distinguish between external and internal validation. The model's ability to generalise to patients from other institutions, geographical locations, imaging devices, or clinical workflows cannot be determined by randomly dividing data from a single dataset, but it may demonstrate performance under comparable data settings. Overly optimistic estimations of model performance may also be caused by potential patient overlap, improper image-level splitting, or other types of information leakage. As a result, the data-splitting approach and test population independence should be considered when interpreting reported accuracy and other performance indicators. For clinical translation, external validation is still crucial. Stronger proof of generalisability and performance degradation brought on by distribution shifts can be found through independent validation employing cohorts that are geographically and institutionally unique. Thus, there is a significant methodological vacuum in the examined literature due to the sparse use of external validation. Multicenter datasets, patient-level separation between training and test data, independent external cohorts, and open documentation of dataset building and validation processes should be given top priority in future research.

Bibliometric and Methodological Insights

In revealing the history of methods and locations of research in the area, the bibliometric analysis goes further than the description of the level of publication. The historical publication trend indicates the transition from the initial emphasis on rapid construction of models and their optimisation to such general problems as clinical use, robustness, interpretability, and validation. It is an illustration of the increasing recognition of the fact that successful medical AI requires more than good performance in experimental conditions.

The geographical concentration of research contributions is further illustrated in Figure 10 below, where China and India contribute significantly to the number of papers reviewed. The population of patients, imaging techniques used, healthcare institutions, and characteristics of the dataset are all influenced by this geographical concentration. It becomes increasingly necessary, therefore, to find out whether models constructed in specific settings will generalise to other populations and health systems through international cooperation and diversified datasets.

How to cite this article:

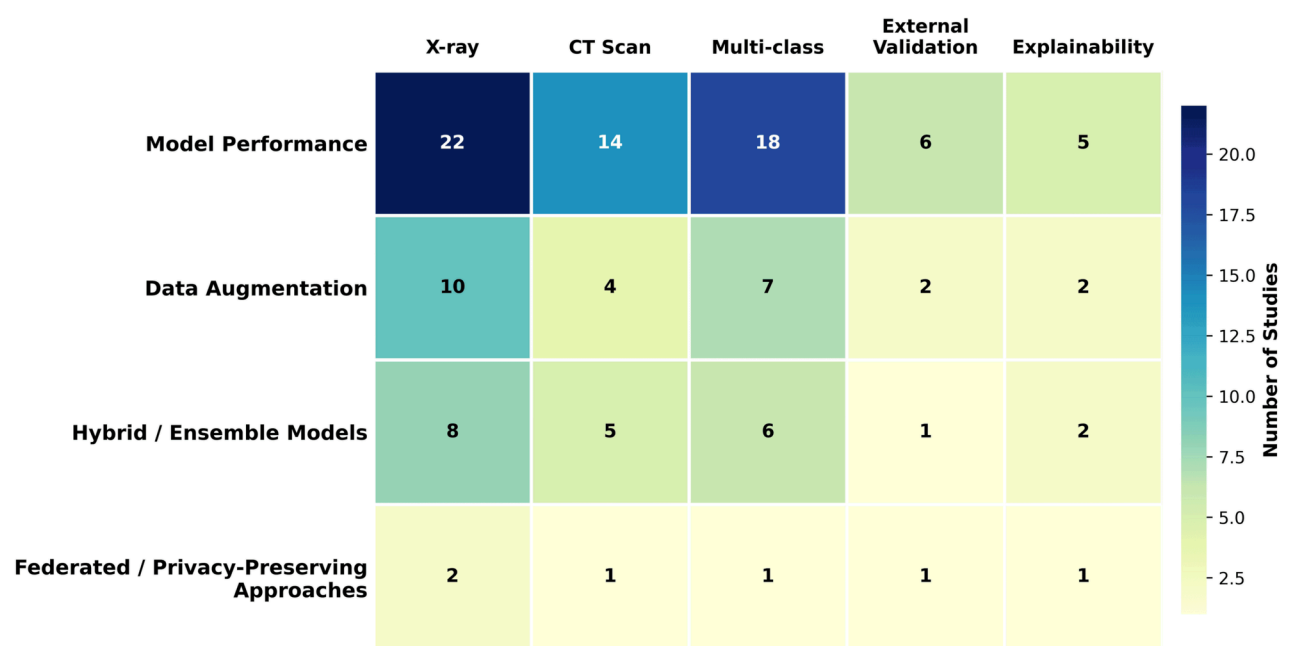


FIGURE 11: The Distribution of Important Study Features Across Various Imaging Modalities in the Evaluated COVID-19 Image Classification Literature Is Depicted in a Research Coverage Matrix. The Number of Studies Is Represented by Cell Values, and the Relative Concentration of Research Activity Is Shown by Colour Intensity

COVID-19, Coronavirus Disease; CT, Computed Tomography

Further evidence of an imbalance between traditional performance evaluation and more general methodological concerns may be seen in the study coverage matrix shown in Figure 17. Across the studied qualities, model performance is significantly more represented than explainability, federated or privacy-preserving techniques, and external validation. This pattern suggests that while very few studies address the needs required to ensure robustness, interpretability, privacy, and translational readiness, research interest is still heavily focused on establishing predictive performance. Crucially, since a single study may address more than one methodological feature, the matrix should be read as a representation of research coverage rather than as proof that these traits are mutually incompatible.

Implications for Clinical Translation and Future Research

When viewed together, the findings imply that there is a mismatch between clinical translation and technological success. Literature indicates that it is possible to discover valuable patterns in COVID-19 medical images using deep learning; however, whether the models will be useful clinically will depend upon their ability to succeed in different clinical scenarios and among varying demographics. Accuracy achieved by the models in one publicly available dataset is no proof of their safety and applicability in treatment. Consequently, depending on the clinical task at hand, future studies should focus on more than simply accuracy-based assessment criteria. Such assessment criteria should comprise sensitivity, specificity, AUC, F1-score, calibration, uncertainty quantification, interpretability, and external validation. Collaboration across multicentre and multi-institutional levels, geographic variation of the datasets, repeatability, and transparency regarding data pretreatment and validation processes deserve more attention. In cases where direct sharing of patients' medical images is forbidden, privacy-preserving methods such as federated learning could present some opportunities for collaborative model development. Moreover, in a related manner, methods for explainable AI could help doctors understand the basis of the model predictions, while uncertainty-aware systems might aid in pointing out situations requiring further expert analysis. Nevertheless, instead of being seen as inherently reliable because of their complex methodology, such methods should be judged based on complete independent verification. In any case, all in all, the scientific field is increasingly moving beyond the stage of fast development of algorithms for emergency cases to

How to cite this article:

a more advanced stage of robustness, generalisability, interpretability, replicability, and clinical applicability. The central implication of this review is therefore that future progress should not be measured only by increasingly high classification accuracy or increasingly complex architectures. Instead, meaningful advancement will depend on demonstrating that deep learning systems can provide transparent, reliable, reproducible, and clinically meaningful decision support across diverse real-world settings and, where appropriate, future infectious disease outbreaks.

The present review found that deep learning-based approaches to automated classification of medical images related to COVID-19 are reporting encouraging results, with many studies achieving over 95% classification accuracy. However, the bibliometric and methodological analyses demonstrated that the development of the field should not be gauged only by quantitative metrics such as the sheer number of published articles and their accuracy scores. The temporal trends identified in the present study demonstrate a paradigm shift, from a focus on model development and optimisation to an emphasis on explainability, robustness, external validation, and other factors. This shift reflects a growing recognition that high levels of accuracy in controlled experiments do not equal generalisability and robustness. In particular, accuracy as a metric is insufficient for evaluating the performance of diagnostic models, as it depends strongly on prevalence. A more informative metric (or metrics) for such cases is a combination of sensitivity and specificity along with an analysis of other metrics, such as AUC, precision, recall, and F1-score, as well as the choice of an appropriate validation strategy. For example, for screening tests, very high sensitivity is critical, as a negative result is taken as a signal to stop further testing; thus, a false-negative result (FN) would entail wasted resources and, possibly, continued spread of the disease. On the other hand, a false-positive result (FP) would lead to extra testing and potentially harmful treatment, extra expenses, and increased stress for the patients. Thus, for a diagnostic test, there is a need to balance specificity and sensitivity in accordance with the particular circumstances to ensure that FPs and FNs are minimised. The present review also identified that external validation, explainability, calibration, and uncertainty quantification are under-researched areas compared to model performance evaluation. This conclusion can be drawn from Figure 11, which demonstrates that most of the studies focus on model performance, whereas external validation, explainability, and privacy-preserving approaches receive comparatively less attention. These topics are of particular interest because most studies report very high levels of accuracy, which can be driven by optimistic within-sample estimates. Thus, methods that allow distinguishing between certain and uncertain predictions are critical for clinical settings, in which incorrect predictions can be costly. Additionally, calibration and uncertainty estimates are important for ensuring that predicted probabilities are not misleading; if they are incorporated into decision-making software, such flawed probabilities might hinder the doctors' own judgements and lead to incorrect conclusions. This is especially true for settings where models are trained on non-representative data (e.g., with few population groups) and/or evaluated internally, and thus, information leakage from the test set into the training data and confirmation bias might inflate performance metrics unrealistically.

The bibliometric analysis identified several trends related to the application of deep learning to COVID-19 medical image classification tasks. First, most of the studies analysed in the present article originate from a small number of countries, which raises questions about the representativeness of the samples and data used in these studies. This limitation is especially evident for studies originating from the US, the UK, or China. Thus, further international collaboration, large-scale studies with diverse samples, and independent validation of widely used algorithms are needed to identify whether and how these approaches generalise to other populations. Secondly, the methodological reviews demonstrate that the field has been progressing rapidly, with researchers exploring new deep learning architectures, addressing multimodality, implementing methods for enhancing privacy, and adopting novel approaches related to explainable AI. Thus, the focus of the studies identified in the present review has shifted from simply demonstrating that deep learning can be used for classification tasks to developing more comprehensive and reliable medical image analysis solutions that consider additional sources of information. The insights from the present review can serve as a foundation for future research in the area of deep learning-based classification of medical images related to COVID-19. In particular, future studies should build upon the findings identified in this review. The findings suggest that researchers should move beyond accuracy and consider other evaluation metrics, such as sensitivity and specificity. Moreover, the selection of an appropriate validation strategy, along with multicentre, multi-institutional collaboration, is paramount. The present review also suggests that further research should focus on developing and validating diagnostic models in external populations and addressing issues of generalisability and reproducibility. Lastly, calibration and incorporation of

How to cite this article:

uncertainty estimation should be considered when developing deep learning-based diagnostic software. At that point, the field will be able to assist medical decision-making in a transparent, accountable, and trustworthy manner rather than only detecting patterns (classification).

Conclusions

To diagnose COVID-19, this study critically examined deep learning-based image classification techniques, concentrating on model architectures, data practices, assessment techniques, and translational preparedness. According to the synthesis, models based on convolutional neural networks - particularly those that make use of transfer learning, ensemble methods, and hybrid architectures - have consistently demonstrated diagnostic performance above 95% on CT and chest X-ray images. Methodological issues still exist despite the encouraging degree of excellent diagnostic performance. "Robustness in clinical practice is limited, and confidence in the reported performance level is eroded by dataset sizes that are small and imbalanced, low levels of external validation, inconsistent sample preparation techniques, and a lack of emphasis on interpretability. Therefore, high precision in experimental performance alone does not necessarily translate to diagnostic efficacy."

Future work must focus on robustness, generalisability, and relevance for treatment rather than optimisation for performance. Multicentre datasets, external validation of models prospectively, and single- versus multiple-disease evaluation of metrics based on clinical risk as opposed to accuracy need further study. Deployments that can be trusted require privacy-preserving and governance-aware systems, whereas interpretability, uncertainty estimation, and transparency must be key design principles. By addressing these requirements, deep learning-based image classification systems might develop from emergency-driven research artefacts into reliable decision support tools that enhance clinical workflow and sustainably support AI-enabled healthcare systems.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Sagardeep Ghosh, Yogendra Chhetri, Pranay Adak, Pijush Dutta

Acquisition, analysis, or interpretation of data: Sagardeep Ghosh, Yogendra Chhetri, Kakali Das, Amrita Mitra

Drafting of the manuscript: Sagardeep Ghosh

Critical review of the manuscript for important intellectual content: Sagardeep Ghosh, Yogendra Chhetri, Kakali Das, Amrita Mitra, Pranay Adak, Pijush Dutta

Supervision: Pranay Adak, Pijush Dutta

Disclosures

Conflicts of interest: In compliance with [the ICMJE uniform disclosure form](#), all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

How to cite this article:

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. Irede EL, Aworinde OR, Ogunnaike KL, Amienghemhen OD, Okonkwo TP, Onivefu AP, Ifijen IH: [Medical imaging: A critical review on X-ray imaging for the detection of infection](#). Biomedical Materials & Devices. 2026, 4:1-45. [10.1007/s44174-024-00212-1](#)
2. Asif S, Yi W, Ur-Rehman S, et al.: [Advancements and prospects of machine learning in medical diagnostics: Unveiling the future of diagnostic precision](#). Archives of Computational Methods in Engineering. 2025, 32:853-883. [10.1007/s11831-024-10148-w](#)
3. Ker JK, Wang L, Rao J, Lim T: [Deep learning applications in medical image analysis](#). IEEE Access. 2018, 6:9375-9389. [10.1109/access.2017.2788044](#)
4. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR: [Automated detection of COVID-19 cases using deep neural networks with X-ray images](#). Computers in Biology and Medicine. 2020, 121:103792.
5. Shuja J, Alanazi E, Alasmay W, Alashaikh A: [COVID-19 open source data sets: A comprehensive survey](#). Applied Intelligence. 2021, 51:1296-1325. [10.1101/2020.05.19.20107532](#)
6. Topol EJ: [High-performance medicine: the convergence of human and artificial intelligence](#). Nature Medicine. 2019, 25:44-56. [10.1038/s41591-018-0300-7](#)
7. Esmaeilzadeh P: [Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives](#). BMC Medical Informatics and Decision Making. 2020, 20:170. [10.1186/s12911-020-01191-1](#)
8. Bhosale YH, Patnaik KS: [Application of deep learning techniques in diagnosis of COVID-19 \(Coronavirus\): A systematic review](#). Neural Processing Letters. 2022, 55:3551-3603. [10.1007/s11063-022-11023-0](#)
9. Ali M, Naeem F, Tariq M, Kaddoum G: [Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey](#). IEEE Journal of Biomedical and Health Informatics. 2023, 27:778-789. [10.1109/jbhi.2022.3181823](#)
10. Torkamaan H, Steinert S, Pera MS, et al.: [Challenges and future directions for integration of large language models into socio-technical systems](#). Behaviour & Information Technology. 2024, 1-20. [10.1080/0144929x.2024.2431068](#)
11. Murphy J, Bustin SA: [Reliability of real-time reverse-transcription PCR in clinical diagnostics: gold standard or substandard?](#). Expert Review of Molecular Diagnostics. 2009, 9:187-197. [10.1586/14737159.9.2.187](#)
12. Al Nufaiei ZF, Alshamrani KM: [Comparing ultrasound, chest X-ray, and CT scan for pneumonia detection](#). Medical Devices: Evidence and Research. 2025, 2025:149-159. [10.2147/mder.s501714](#)
13. Ye J, Ni J, Yi Y: [Deep learning hierarchical representations for image steganalysis](#). IEEE Transactions on Information Forensics and Security. 2017, 12:2545-2557. [10.1109/tifs.2017.2710946](#)
14. Alyasseri ZAA, Al-Betar MA, Doush IA, et al.: [Review on COVID-19 diagnosis models based on machine learning and deep learning approaches](#). Expert Systems. 2021, 39:e12759. [10.1111/exsy.12759](#)
15. Hariri W, Narin A: [Deep neural networks for COVID-19 detection and diagnosis using images and acoustic-based techniques: a recent review](#). Soft Computing. 2021, 25:15345-15362. [10.1007/s00500-021-06137-x](#)
16. Jaiswal A, Gianchandani N, Singh D, Kumar V, Kaur M: [Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning](#). Journal of Biomolecular Structure and Dynamics. 2021, 39:5682-5689. [10.1080/07391102.2020.1788642](#)
17. Ismael AM, Şengür A: [Deep learning approaches for COVID-19 detection based on chest X-ray images](#). Expert Systems with Applications. 2021, 164:114054. [10.1016/j.eswa.2020.114054](#)
18. Sethy PK, Behera SK, Ratha PK, Biswas P: [Detection of coronavirus disease \(COVID-19\) based on deep features and support vector machine](#). International Journal of Mathematical, Engineering and Management Sciences. 2020, 5:643-651. [10.33889/ijmems.2020.5.4.052](#)
19. Ibrahim AU, Ozsoz M, Serte S, Al-Turjman F, Yakoi PS: [Pneumonia classification using deep learning from chest X-ray images during COVID-19](#). Cognitive Computation. 2024, 16:1589-1601. [10.1007/s12559-020-09787-5](#)
20. Amyar A, Modzelewski R, Li H, Ruan S: [Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation](#). Computers in Biology and Medicine. 2020, 126:104037. [10.1016/j.compbiomed.2020.104037](#)

How to cite this article:

21. Zhou T, Lu H, Yang Z, Qiu S, Huo B, Dong Y: [The ensemble deep learning model for novel COVID-19 on CT images](#). Applied Soft Computing. 2021, 98:106885. [10.1016/j.asoc.2020.106885](#)
22. Ravi V, Narasimhan H, Chakraborty C, Pham TD: [Deep learning-based meta-classifier approach for COVID-19 classification using CT scan and chest X-ray images](#). Multimedia Systems. 2022, 28:1401-1415. [10.1007/s00530-021-00826-1](#)
23. Kanwal A, Chandrasekaran S: [2D CNN-BiCuDNNLSTM: hybrid deep-learning-based approach for classification of COVID-19 X-ray images](#). Sustainability. 2022, 14:6785. [10.3390/su14116785](#)
24. Aggarwal P, Mishra NK, Fatimah B, Singh P, Gupta A, Joshi SD: [COVID-19 image classification using deep learning: Advances, challenges and opportunities](#). Computers in Biology and Medicine. 2022, 144:105350. [10.1016/j.combiomed.2022.105350](#)
25. Li X, Tan W, Liu P, Zhou Q, Yang J: [Classification of COVID-19 chest CT images based on ensemble deep learning](#). Journal of Healthcare Engineering. 2021, 2021:5528441. [10.1155/2021/5528441](#)
26. Malik H, Naeem AA, Naqvi RA, Loh WK: [DMFL_Net: A federated learning-based framework for the classification of COVID-19 from multiple chest diseases using X-rays](#). Sensors. 2023, 23:743. [10.3390/s23020743](#)
27. Pathak Y, Shukla PK, Tiwari A, Stalin S, Singh S, Shukla PK: [Deep transfer learning based classification model for COVID-19 disease](#). Innovation and Research in BioMedical Engineering. 2022, 43:87-92. [10.1016/j.irbm.2020.05.003](#)
28. Kassania SH, Kassanib PH, Wesolowskic MJ, Schneidera KA, Detersa R: [Automatic detection of coronavirus disease \(COVID-19\) in X-ray and CT images: A machine learning based approach](#). Biocybernetics and Biomedical Engineering. 2021, 41:867-879. [10.1016/j.bbe.2021.05.013](#)
29. Ahmadian S, Jalali SMJ, Islam SMS, Khosravi A, Fazli E, Nahavandi S: [A novel deep neuroevolution-based image classification method to diagnose coronavirus disease \(COVID-19\)](#). Computers in Biology and Medicine. 2021, 139:104994. [10.1016/j.combiomed.2021.104994](#)
30. Kamal KC, Yin Z, Wu M, Wu Z: [Evaluation of deep learning-based approaches for COVID-19 classification based on chest X-ray images](#). Signal, Image and Video Processing. 2021, 15:959-966. [10.1007/s11760-020-01820-2](#)
31. Hussain E, Hasan M, Rahman MA, Lee I, Tamanna T, Parvez MZ: [CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images](#). Chaos, Solitons & Fractals. 2021, 142:110495. [10.1016/j.chaos.2020.110495](#)
32. Sakib S, Tazrin T, Fouda MM, Fadlullah ZM, Guizani M: [DL-CRC: Deep learning-based chest radiograph classification for COVID-19 detection: A novel approach](#). IEEE Access. 2020, 8:171575-171589. [10.1109/access.2020.3025010](#)
33. Yoo SH, Geng H, Chiu TL, et al.: [Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging](#). Frontiers in Medicine. 2020, 7:427. [10.3389/fmed.2020.00427](#)
34. Surya MBR, Rao YM, Venkataiah C, Murthy GLN, Dharani M, Jayamma M: [Deep learning based classification of COVID-19 severity using hierarchical deep maxout model](#). Biomedical Signal Processing and Control. 2024, 88:105653. [10.1016/j.bspc.2023.105653](#)
35. Thilagavathi P, Geetha R, Jothi Shri S, Somasundaram K: [An effective COVID-19 classification in X-ray images using a new deep learning framework](#). Journal of X-Ray Science and Technology. 2025, 33:297-316. [10.1177/08953996241290893](#)
36. Appasami G, Nickolas S: [A deep learning-based COVID-19 classification from chest X-ray image: case study](#). The European Physical Journal Special Topics. 2022, 231:3767-3777. [10.1140/epjs/s11734-022-00647-x](#)
37. Abdel-Basset M, Chang V, Hawash H, Chakraborty RK, Ryan M: [FSS-2019-nCov: A deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection](#). Knowledge-Based Systems. 2021, 212:106647. [10.1016/j.knosys.2020.106647](#)
38. Kong L, Cheng J: [Classification and detection of COVID-19 X-Ray images based on DenseNet and VGG16 feature fusion](#). Biomedical Signal Processing and Control. 2022, 77:103772. [10.1016/j.bspc.2022.103772](#)
39. Kumar S, Bhowmik B: [EffiCOVID-net: A highly efficient convolutional neural network for COVID-19 diagnosis using chest X-ray imaging](#). Methods. 2025, 240:81-100. [10.1016/j.ymeth.2025.04.008](#)
40. Narin A, Kaya C, Pamuk Z: [Automatic detection of coronavirus disease \(COVID-19\) using X-ray images and deep convolutional neural networks](#). Pattern Analysis and Applications. 2021, 24:1207-1220. [10.1007/s10044-021-00984-y](#)
41. Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C: [Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound](#). IEEE Transactions on Medical Imaging. 2020, 39:2676-2687. [10.1109/tmi.2020.2994459](#)
42. Ouchicha C, Ammor O, Meknassi M: [CVDNet: A novel deep learning architecture for detection of coronavirus \(Covid-19\) from chest x-ray images](#). Chaos, Solitons & Fractals. 2020, 140:110245. [10.1016/j.chaos.2020.110245](#)

How to cite this article:

43. Barshooi AH, Amirkhani A: [A novel data augmentation based on Gabor filter and convolutional deep learning for improving the classification of COVID-19 chest X-Ray images](#). Biomedical Signal Processing and Control. 2022, 72:103326. [10.1016/j.bspc.2021.103326](#)
44. Dhiman G, Chang V, Kant SK, Shankar A: [ADOPT: automatic deep learning and optimization-based approach for detection of novel coronavirus COVID-19 disease using X-ray images](#). Journal of Biomolecular Structure and Dynamics. 2022, 40:5836-5847. [10.1080/07391102.2021.1875049](#)
45. Zhao W, Jiang W, Qiu X: [Deep learning for COVID-19 detection based on CT images](#). Scientific Reports. 2021, 11:14353. [10.1038/s41598-021-93832-2](#)
46. Ibrahim DM, Elshennawy NM, Sarhan AM: [Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases](#). Computers in Biology and Medicine. 2021, 132:104348. [10.1016/j.combiomed.2021.104348](#)
47. Yang D, Martinez C, Visuña L, Khandhar H, Bhatt C, Carretero J: [Detection and analysis of COVID-19 in medical images using deep learning techniques](#). Scientific Reports. 2021, 11:19638. [10.1038/s41598-021-99015-3](#)
48. Khan AI, Shah JL, Bhat MM: [CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images](#). Computer Methods and Programs in Biomedicine. 2020, 196:105581. [10.1016/j.cmpb.2020.105581](#)
49. Tian G, Wang Z, Wang C, et al.: [A deep ensemble learning-based automated detection of COVID-19 using lung CT images and Vision Transformer and ConvNeXt](#). Frontiers in Microbiology. 2022, 13:1024104. [10.3389/fmicb.2022.1024104](#)
50. Bhargava A, Bansal A, Goyal V: [Machine learning-based automatic detection of novel coronavirus \(COVID-19\) disease](#). Multimedia Tools and Applications. 2022, 81:13731-13750. [10.1007/s11042-022-12508-9](#)
51. Jain G, Mittal D, Thakur D, Mittal MK: [A deep learning approach to detect Covid-19 coronavirus with X-ray images](#). Biocybernetics and Biomedical Engineering. 2020, 40:1391-1405. [10.1016/j.bbe.2020.08.008](#)
52. Xu Y, Lam HK, Jia G: [MANet: A two-stage deep learning method for classification of COVID-19 from chest X-ray images](#). Neurocomputing. 2021, 443:96-105. [10.1016/j.neucom.2021.03.034](#)
53. Gaur L, Bhatia U, Jhanjhi NZ, Muhammad G, Masud M: [Medical image-based detection of COVID-19 using deep convolution neural networks](#). Multimedia Systems. 2023, 29:1729-1738. [10.1007/s00530-021-00794-6](#)
54. Joshi RC, Yadav S, Pathak VK, et al.: [A deep learning-based COVID-19 automatic diagnostic framework using chest X-ray images](#). Biocybernetics and Biomedical Engineering. 2021, 41:239-254. [10.1016/j.bbe.2021.01.002](#)
55. Loey M, El-Sappagh S, Mirjalili S: [Bayesian-based optimized deep learning model to detect COVID-19 patients using chest X-ray image data](#). Computers in Biology and Medicine. 2022, 142:105213. [10.1016/j.combiomed.2022.105213](#)
56. Akter S, Shamrat FMJM, Chakraborty S, Karim A, Azam S: [COVID-19 detection using deep learning algorithm on chest X-ray images](#). Biology. 2021, 10:1174. [10.3390/biology10111174](#)
57. Goldstein E, Keidar D, Yaron D, et al.: [COVID-19 classification of X-ray images using deep neural networks](#). arXiv. 2020, [10.48550/arXiv.2010.01362](#)
58. Bhattacharyya A, Bhaik D, Kumar S, Thakur P, Sharma R, Pachori RB: [A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images](#). Biomedical Signal Processing and Control. 2022, 71:103182. [10.1016/j.bspc.2021.103182](#)
59. Horry MJ, Chakraborty S, Paul M, Ulhaq A, Pradhan B, Saha M, Shukla N: [COVID-19 detection through transfer learning using multimodal imaging data](#). IEEE Access. 2020, 8:149808-149824. [10.1109/access.2020.3016780](#)
60. Mohammadi R, Salehi M, Ghaffari H, Rohani AA, Reiazi R: [Transfer learning-based automatic detection of coronavirus disease 2019 \(COVID-19\) from chest X-ray images](#). Journal of Biomedical Physics and Engineering. 2020, 10:559-568. [10.31661/jbpe.v0i0.2008-1153](#)

How to cite this article: