IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A REVIEW OF DEEP LEARNING AND CNN ENSEMBLE METHODS FOR COVID-19 DETECTION FROM CHEST X-RAY IMAGES

Miss. Sudke Isha Rajan[1], Dr. Damre S.S.[2]

[1]PG Student, M.S. Bidve Engineering College, Latur, Maharashtra, India

[2]Professor, M.S. Bidve Engineering College, Latur, Maharashtra, India

Abstract: The COVID-19 pandemic created sustained demand for rapid diagnostic triage, and chest radiography was widely investigated as a fast, inexpensive and broadly available complement to reverse transcription polymerase chain reaction testing. A large body of work published from 2020 onwards applied deep learning, and convolutional neural networks (CNNs) in particular, to the automated interpretation of chest radiographs for this purpose. This paper reviews that literature. It examines the CNN architectures most frequently adopted, including VGG, ResNet, DenseNet, Inception, Xception, MobileNet and EfficientNet; the near-universal reliance on transfer learning from ImageNet; the public datasets on which the field has depended and the preprocessing and augmentation pipelines applied to them; the ensemble and hybrid strategies proposed to reduce dependence on any single architecture; and the explainability methods, chiefly Grad-CAM, used to expose the evidence behind a prediction. Reported accuracies across the reviewed studies are frequently high, but the review finds that these figures are not directly comparable, because studies differ in dataset composition, class definitions, partitioning, preprocessing and evaluation protocol. Critical appraisals of the field have further shown that models can separate classes using source-specific artefacts rather than pathology, and that performance degrades substantially under cross-dataset evaluation. The review identifies persistent challenges, including small and imbalanced datasets, dataset-source bias, data leakage, the scarcity of external validation and the absence of standardised benchmarks, and outlines the research directions the literature indicates are needed.

Index Terms - COVID-19, Chest X-ray, Convolutional Neural Network, Deep Learning, Transfer Learning, Ensemble Learning, Literature Review, Explainable Artificial Intelligence, Grad-CAM, Medical Image Classification, Dataset Bias

I. INTRODUCTION

The coronavirus disease 2019 (COVID-19) outbreak, caused by SARS-CoV-2, developed into a global public health emergency within months of its identification in late 2019. Health systems were placed under acute strain, and the capacity to identify infected individuals rapidly became central to isolation, triage and treatment planning. Diagnostic throughput, rather than diagnostic possibility, was frequently the binding constraint.

Reverse transcription polymerase chain reaction (RT-PCR) testing of nasopharyngeal swabs is the accepted reference standard. It is highly specific, but requires laboratory infrastructure and trained personnel, and its sensitivity depends on sampling quality and on the stage of infection at which the specimen is taken [11]. These properties motivated interest in imaging as a complementary triage modality. Chest radiography is inexpensive, widely distributed, portable to the bedside, and rapid to acquire, and COVID-19 pneumonia is associated with characteristic radiographic appearances [5], [9]. Radiographic findings are not specific to COVID-19 and overlap with those of other viral and bacterial pneumonias, so imaging has generally been framed in the literature as a screening and triage aid rather than a substitute for molecular confirmation.

Interpreting radiographs at scale requires experienced readers whose availability is uneven, which prompted extensive work on automated interpretation. Convolutional neural networks (CNNs) learn hierarchical representations directly from pixel data and have displaced handcrafted feature engineering across most medical image analysis tasks [21], [27]. Because the COVID-19 radiograph collections assembled during the pandemic were small, almost all reported work adopted transfer learning, initialising deep backbones such as ResNet50 [1], DenseNet121 [2] and VGG19 [3] from ImageNet weights and fine-tuning them on the target data [22].

Two further themes recur throughout this literature. The first is ensemble learning. Combining predictors whose errors are imperfectly correlated reduces variance and typically improves generalisation relative to the individual members [23], [24], [29], and numerous studies have combined several CNN backbones by voting, averaging, weighted averaging, stacking or feature concatenation. The second is explainability. A classifier that outputs only a label gives a clinician no means of assessing whether the decision rests on lung parenchyma or on an incidental artefact, and gradient-weighted class activation mapping (Grad-CAM) [4], building on class activation mapping [25], has become the default method for producing coarse localisation maps in this domain.

A review of this literature is warranted for several reasons. The volume of publication was unusually high and unusually rapid, which compressed the normal cycle of scrutiny. Reported accuracies are frequently high and superficially similar, yet the underlying studies differ so substantially in dataset composition, class definitions, partitioning and evaluation protocol that the numbers are not directly comparable. A subsequent body of critical work has questioned whether the reported performance reflects genuine learning of pathology at all [19], [20], [31]. Synthesising these strands, rather than enumerating individual results, is necessary to establish what the field has actually demonstrated.

The objective of this review is to examine existing deep learning approaches for COVID-19 detection from chest X-ray images, with particular emphasis on CNN architectures, transfer learning, ensemble strategies, datasets and preprocessing techniques, and explainable artificial intelligence. The review compares and synthesises the reported findings, examines why performance varies across studies, and identifies the challenges and research gaps that the literature itself has surfaced.

The remainder of the paper is organised as follows. Section II provides background on COVID-19 and chest radiographic imaging. Section III reviews the deep learning methods applied to the task. Section IV examines the datasets and preprocessing pipelines on which the field depends. Section V surveys single-model, transfer-learning, hybrid and ensemble detection methods. Section VI reviews explainable AI methods. Section VII presents a comparative analysis across studies. Section VIII sets out the challenges and research gaps, Section IX the future directions indicated by the literature, and Section X concludes.

II. BACKGROUND OF COVID-19 AND CHEST X-RAY IMAGING

A. Diagnostic Context

SARS-CoV-2 infection is confirmed by RT-PCR, which detects viral RNA. The literature on imaging-based screening consistently frames its motivation in terms of the practical limitations of that reference standard rather than its accuracy: turnaround time, dependence on laboratory capacity and reagent supply, and reduced sensitivity in early infection [11]. Where laboratory capacity is constrained, the question addressed by imaging is not whether a radiograph can replace a molecular test, but whether it can help decide which patients are tested and isolated first.

B. Radiographic Findings Reported in COVID-19

The radiographic appearances associated with COVID-19 pneumonia are described across the reviewed studies as bilateral, predominantly peripheral and lower-zone ground-glass opacification and consolidation [5], [9]. These appearances are not pathognomonic. They overlap substantially with those of other viral pneumonias and, in the early phase of infection, radiographs may appear normal. This overlap is the reason that most reviewed studies frame the task as classification among COVID-19, other pneumonia and normal categories rather than as diagnosis, and it also explains why three-class formulations are consistently reported as harder than binary ones.

C. Chest Radiography Compared with Computed Tomography

Computed tomography is more sensitive to the parenchymal changes of COVID-19 than plain radiography, but is more expensive, less widely available, delivers a higher radiation dose and requires scanner decontamination between infectious patients. The reviewed literature on radiography justifies its focus on these logistical grounds. Horry et al. [17] compared X-ray, CT and ultrasound within a single transfer

learning study, and their work illustrates both the appeal of cross-modality comparison and the difficulty of drawing conclusions from the small COVID-19 sample sizes available in each modality.

III. DEEP LEARNING FOR COVID-19 DETECTION

This section reviews the network architectures and learning strategies reported across the literature. Figure 1 organises the categories of approach encountered.

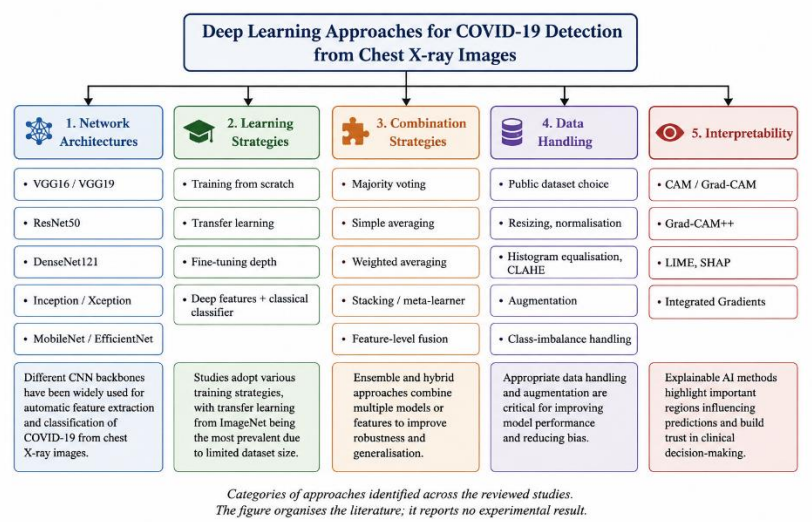

Fig. 1. Taxonomy of deep learning approaches for COVID-19 detection from chest X-ray images reported in the reviewed literature.

A. Convolutional Neural Networks

A CNN applies learned convolution kernels across the image, so that the same feature detector is evaluated at every spatial position. Weight sharing and local connectivity reduce the parameter count relative to a fully connected network and introduce translation equivariance, while stacking convolution and pooling layers builds progressively larger receptive fields, so that early layers respond to edges and texture and later layers to larger structures [21], [27]. In the reviewed studies the final convolutional feature map is pooled and passed to a fully connected layer with a softmax activation producing class posteriors. This formulation is common to essentially all the classification studies reviewed here; the differences between them lie in the architecture producing the class scores, in how that architecture is trained, and in the data on which it is trained. Those three axes, rather than the classification layer itself, account for the variation in published results examined in Section VII.

B. Transfer Learning

Transfer learning reuses parameters learned on a large source dataset, in practice almost always ImageNet [21], as the initialisation for a smaller target task [22]. Its dominance in this literature follows directly from data scarcity: the public COVID-19 radiograph collections contain positive classes numbering in the hundreds, which is far below what networks of this depth require when trained from random initialisation. The justification usually offered is that early convolutional layers encode generic edge, texture and gradient detectors that transfer across image domains, so that only the later, more task-specific layers need substantial adaptation.

Studies differ in how much of the network they adapt. Some freeze the convolutional trunk entirely and train only a replaced classification head; some unfreeze the final blocks; some fine-tune all layers at a reduced learning rate. Others bypass end-to-end training altogether, extracting deep features from a frozen backbone and classifying them with a support vector machine or similar classical model, as reported by Ismael and Sengur [14]. The literature does not establish a single preferred strategy, and because studies vary simultaneously in dataset, preprocessing and fine-tuning depth, the independent effect of that choice is difficult to isolate from published results.

A caveat that recurs in the critical literature is that transfer from natural images to radiographs is a substantial domain shift. Grayscale radiographs have different intensity statistics and spatial structure from ImageNet photographs, and the benefit of ImageNet initialisation in medical imaging has been questioned in the broader literature. Where target datasets are very small, ImageNet initialisation may improve reported performance chiefly by regularising against overfitting rather than by supplying genuinely transferable pathological features.

C. ResNet-Based Methods

The residual network of He et al. [1] introduced identity shortcut connections, so that a block learns a residual function F(x) added to its own input and computes F(x) + x. Two consequences are commonly cited. Gradients propagate to early layers through the shortcut without repeated multiplication by intermediate Jacobians, which mitigates the vanishing-gradient problem and permits stable training at depth; and a block can approximate the identity mapping by driving its residual branch towards zero, so that adding depth does not necessarily degrade the representation. ResNet50, the 50-layer bottleneck variant, is among the most frequently adopted backbones in the reviewed studies. Narin et al. [6] compared it with InceptionV3 and Inception-ResNetV2 and reported it the strongest of the three; Minaee et al. [12] included ResNet18 and ResNet50 among four fine-tuned backbones; and Ismael and Sengur [14] used ResNet50 as a deep feature extractor. Rahimzadeh and Attar [13] used ResNet50V2 as one of two concatenated branches.

D. DenseNet-Based Methods

DenseNet [2] connects each layer within a dense block to every subsequent layer by channel-wise concatenation rather than by summation, so that a layer receives the feature maps of all preceding layers in its block. The properties usually cited are feature reuse, since features computed early are used directly by later layers instead of being reconstructed; strengthened gradient flow, since every layer has short-path access to the loss; and parameter efficiency, since the number of new feature maps per layer can be kept small. DenseNet121 has a well-established precedent in chest radiography independent of COVID-19: Rajpurkar et al. [32] used it for pneumonia detection on ChestX-ray14 [27], which is frequently cited as justification for its adoption in the COVID-19 setting. Minaee et al. [12] and Chowdhury et al. [9] both included DenseNet121 among the backbones evaluated.

E. VGG-Based Methods

The VGG networks of Simonyan and Zisserman [3] are plain sequential stacks of 3 x 3 convolutions with 2 x 2 max pooling, without shortcut or concatenation connections. VGG16 and VGG19 remain widely used in this literature despite predating the architectures above, and Apostolopoulos and Mpesiana [8] reported VGG19 as competitive with newer backbones on their composite dataset, while Horry et al. [17] adopted VGG19 across three imaging modalities and Nishio et al. [16] used VGG16. The architecture is straightforward, its hierarchical feature extraction is easy to reason about, and its dense early layers preserve fine spatial detail. Its principal limitation, consistently noted, is computational: the fully connected layers carry a large parameter count, giving VGG a substantially higher memory and inference cost than DenseNet or MobileNet at comparable or lower depth. This matters for the resource-limited deployment settings that motivate radiograph-based screening.

F. Lightweight CNN Architectures

Lightweight architectures were investigated for deployment where computation is constrained. MobileNet [28] uses depthwise separable convolutions to reduce parameter count and multiply-accumulate operations by a large factor relative to standard convolution, and Apostolopoulos and Mpesiana [8] reported MobileNetV2 among their stronger performers. Xception [33] takes the depthwise separable idea further as a general-purpose architecture and was used as one branch of the concatenated model of Rahimzadeh and Attar [13]. EfficientNet [34] derives a compound scaling rule for depth, width and resolution and reports strong accuracy-per-parameter on ImageNet. Inception [30], [35] processes each stage at multiple kernel scales in parallel. The reviewed literature does not establish a consistent accuracy penalty for the lightweight families in this task; because the positive-class sample sizes are small, differences between backbones frequently fall within the range plausibly attributable to dataset partitioning and initialisation.

Table 1. Comparison of Common CNN Architectures Used for Medical Image Classification

Architecture	Major Feature	Advantages	Limitations	Typical Medical Imaging Application
VGG16 / VGG19 [3]	Plain sequential stack of 3 x 3 convolutions	Simple and easy to interpret; dense early layers preserve fine spatial detail; widely available pretrained weights	Large parameter count in fully connected layers; high memory and inference cost; no shortcut connections, so depth is limited	Transfer-learning baseline for chest radiograph classification [8], [16], [17]
ResNet50 [1]	Identity shortcut connections and residual blocks	Stable training at depth; mitigates vanishing gradients; strong general-purpose backbone	Substantial computation relative to lightweight families; deep representations may overfit very small datasets	General radiograph and CT classification; deep feature extraction [6], [12], [14]
DenseNet121 [2]	Dense channel-wise concatenation within blocks	Feature reuse reduces redundancy; strong gradient flow; compact parameter count for its depth	Concatenation increases memory use during training; sensitive to growth-rate choice	Chest radiograph pathology classification, established by CheXNet [32]
Inception / GoogLeNet [30], [35]	Parallel multi-scale kernels within a module	Captures features at several scales simultaneously; efficient for its accuracy	Architecturally complex; less straightforward to modify or interpret	Multi-class radiograph classification [6]
Xception [33]	Depthwise separable convolutions throughout	Decouples spatial and cross-channel correlation; efficient use of parameters	Benefit over Inception is modest on small datasets	Used as an ensemble branch for COVID-19 radiographs [13]
MobileNet [28]	Depthwise separable convolutions with width multiplier	Low parameter count and inference cost; suitable for constrained hardware	Reduced representational capacity relative to deeper backbones	Screening on resource-limited or mobile deployment [8]
EfficientNet [34]	Compound scaling of depth, width and input resolution	Strong reported accuracy per parameter; family spans a wide compute range	Sensitive to input resolution; scaling rule derived on natural images, not radiographs	Increasingly reported as a backbone in later radiograph studies

IV. DATASETS AND PREPROCESSING METHODS

Figure 2 shows the processing sequence commonly reported across the reviewed studies. Individual studies omit or reorder stages, and the schematic should be read as an organising device rather than as a specification.

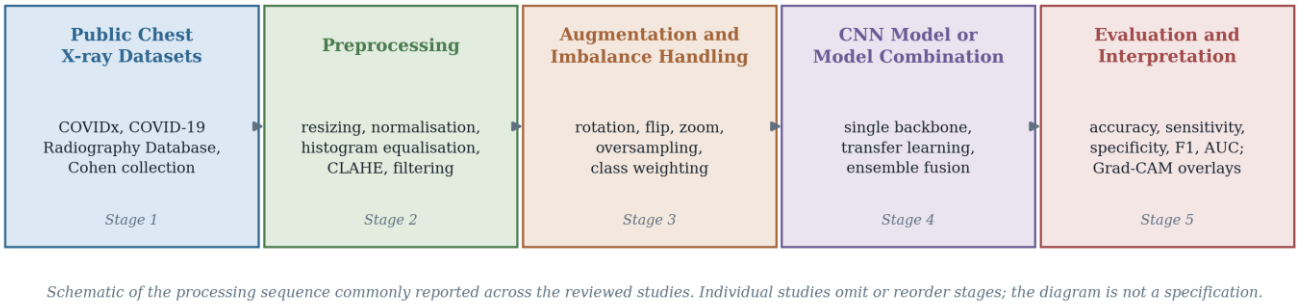

Fig. 2. Generic processing pipeline reported across the reviewed studies.

A. Public Chest X-Ray Datasets

The field has depended on a small number of public collections, several of which are aggregations of other sources. The COVID-19 image data collection of Cohen et al. [11] gathered COVID-19 and other pneumonia radiographs and computed tomography images from published case reports and online sources; it is the origin of the positive class in a large fraction of subsequent work. The COVIDx dataset assembled by Wang et al. [5] alongside COVID-Net combines several open sources into a three-class corpus. The COVID-19 Radiography Database of Chowdhury et al. [9], later extended by Rahman et al. [10], is a widely used curated collection that has been released in successive versions of increasing size. ChestX-ray8 and its extension ChestX-ray14 [27] supply non-COVID comparison images at hospital scale, and CheXpert [36] is a comparable large-scale labelled corpus.

Two properties of these collections are important when reading the reported results. First, the COVID-19 positive class is small relative to the negative classes in almost every version, so class imbalance is intrinsic rather than incidental. Second, and more consequentially, in several widely used composites the positive and negative classes originate from different institutions, imaging equipment and postprocessing pipelines, because the COVID-19 images were drawn from case reports while the normal and pneumonia images came from pre-existing hospital datasets. Section VIII returns to the consequences of this.

Dataset versioning is a persistent source of confusion in this literature. The COVID-19 Radiography Database in particular has been released in successive versions with substantially different numbers of COVID-19 images, and papers citing it by name alone are frequently not describing the same corpus. Similarly, the Cohen collection [11] is a continuously updated repository rather than a fixed release, so two studies citing it may have downloaded materially different data. Where the reviewed studies report dataset sizes, those figures are referred to in Table 4, but readers comparing across rows should treat any difference in reported dataset size as potentially reflecting dataset version rather than deliberate selection.

B. Image Resizing and Normalization

Radiographs are acquired at varying resolution and are resampled to a fixed input size, most commonly 224 x 224 pixels, because that is the input dimension of the ImageNet-pretrained backbones the field relies on. Resizing enables batching and keeps the receptive-field geometry of the pretrained filters consistent with their training conditions, at the cost of discarding spatial detail. Intensity normalisation, typically standardisation to zero mean and unit variance or rescaling to a fixed interval, is applied so that images acquired at different exposures contribute comparably during optimisation. Both operations are near-universal across the reviewed studies and are rarely varied experimentally, so the literature offers little direct evidence on their individual contribution.

C. Histogram Equalization and Enhancement

Contrast enhancement redistributes intensities so that parenchymal detail is more clearly expressed. Global histogram equalisation maps intensities according to their cumulative distribution across the whole image, which increases global contrast but can amplify noise in uniform regions and can distort the appearance of already well-exposed areas. Contrast-limited adaptive histogram equalisation (CLAHE) [26]

operates on local tiles and clips the histogram before equalisation, which limits noise amplification and preserves local detail at higher computational cost.

The evidence that these choices matter is direct. Rahman et al. [10] examined several enhancement techniques and reported that the choice measurably changes downstream classification performance. This is an important result for reading the wider literature, because it implies that part of the variation in reported accuracy between studies is attributable to preprocessing rather than to architecture, and most studies neither vary enhancement systematically nor report it in enough detail for the effect to be separated.

D. Data Augmentation

Augmentation expands the effective training set by applying label-preserving transformations, most commonly small-angle rotation, translation, isotropic zoom and horizontal flipping, and sometimes brightness or contrast jitter and random cropping. Rotation and translation model variation in patient positioning and zoom models variation in source-to-detector geometry. Horizontal flipping warrants more care than it usually receives, since it reverses left-right anatomy and is appropriate only where laterality is not diagnostically relevant. Nishio et al. [16] studied combinations of augmentation operations explicitly and reported that the combination chosen affects the result, which again indicates that some of the between-study variance in the literature originates in preprocessing choices.

Lung-field segmentation is a less common but methodologically important preprocessing step. Restricting the network input to the segmented lungs removes text markers, borders and support hardware from the field of view, and is one of the few preprocessing choices with a direct bearing on the shortcut-learning concern discussed in Section VIII.

E. Class Imbalance

Because the COVID-19 class is a small minority in the available corpora, class imbalance affects both training and evaluation. During training, an unweighted objective biases the classifier towards the majority class. Reported mitigations include random oversampling of the minority class, as used by Punn and Agarwal [15], class-weighted loss functions, and augmentation applied preferentially to the minority class. During evaluation, imbalance makes aggregate accuracy misleading, since a classifier that always predicts the majority class can score highly. The consequences are visible in the reviewed results: Rahimzadeh and Attar [13] reported a high overall accuracy alongside a markedly lower score for the COVID-19 class specifically, which illustrates why per-class metrics are necessary and why aggregate accuracy alone is a weak basis for comparison.

V. CNN-BASED COVID-19 DETECTION METHODS

A. Single CNN Models

A first strand designed or adopted a single network for the task. Wang et al. [5] introduced COVID-Net, an architecture developed specifically for COVID-19 chest radiograph classification using a projection-expansion-projection-extension design to control parameter count, released together with the COVIDx dataset. Ozturk et al. [7] proposed DarkCovidNet, based on the DarkNet-19 backbone, evaluated in both binary and three-class settings. Both released code and data, which improved reproducibility relative to much of the surrounding literature; both were nonetheless trained on positive classes numbering in the low hundreds.

B. Transfer-Learning Approaches

The largest group of studies fine-tuned established ImageNet-pretrained backbones. Apostolopoulos and Mpesiana [8] evaluated several such backbones on a composite public dataset in two-class and three-class settings. Narin et al. [6] compared three backbones across four dataset configurations. Chowdhury et al. [9] compared a larger set of networks with and without augmentation on their curated database. Punn and Agarwal [15] combined fine-tuned backbones with explicit imbalance-handling. Nishio et al. [16] focused on augmentation strategy with a VGG16 backbone. Horry et al. [17] extended the comparison across imaging modalities. These studies share a common structure, and their differences lie principally in dataset, preprocessing and fine-tuning depth rather than in method.

C. Hybrid CNN Approaches

Hybrid approaches combine deep feature extraction with a separate classifier or with a second architecture. Ismael and Sengur [14] extracted deep features from pretrained backbones and classified them with a support vector machine, comparing this against end-to-end fine-tuning. Rahimzadeh and Attar [13] concatenated the feature maps of Xception and ResNet50V2 into a single classifier, which is a feature-level rather than

decision-level combination. The rationale offered for hybrid designs is that a classical classifier trained on fixed deep features has far fewer free parameters than a fine-tuned network and may therefore overfit less on small data.

D. Ensemble Learning Approaches

The theoretical basis for ensembles is long established. Hansen and Salamon [23] showed that combining neural networks whose errors are not perfectly correlated reduces expected error, and Dietterich [24] and Polikar [29] set out the statistical, computational and representational reasons why ensembles help. The benefit depends on member diversity, which is why architecturally distinct backbones are preferable to several instances of the same network.

The fusion strategies reported in this literature fall into several families.

- Majority voting, in which each member casts one vote for its predicted class and the plurality wins. It is simple and requires no tuning, but discards the confidence information in the softmax outputs and can be unstable with an even number of members.
- Simple averaging, in which the members' probability vectors are averaged with equal weight. This retains confidence information and introduces no parameters, but treats a weak member as equal to a strong one.
- Weighted averaging, in which each member receives a scalar weight. Writing P for the probability vector of member i and w for its weight, the fused posterior is the convex combination

$$P_{\text{fused}} = \sum_{i=1}^{N} w_i P_i , \quad \sum_{i=1}^{N} w_i = 1, \quad w_i \geq 0 \tag{1}$$

where N is the number of members and C the number of classes. The non-negativity and unit-sum constraints ensure that the fused vector is itself a valid probability vector. In the reviewed studies the weights are most often fixed a priori or set from validation accuracy; the constraint set is frequently left implicit and the procedure for choosing the weights is often described only briefly.

- Stacking, in which a meta-learner is trained on the members' outputs to learn the combination rule. This is more flexible than a fixed rule but introduces additional parameters and requires a further held-out partition, which is costly when data are scarce.
- Feature-level fusion, in which intermediate representations are concatenated before a shared classification head, as in [13]. This allows interaction between members' features but couples them, so the members can no longer be trained or replaced independently.

Decision-level fusion, covering voting and the averaging rules, keeps members independently trainable and reduces each member's contribution to an interpretable scalar. Feature-level fusion is more expressive but sacrifices that modularity. The reviewed literature does not settle which is preferable for this task; the studies adopting each differ in too many other respects for the comparison to be isolated.

Table 2. Existing Ensemble and Hybrid Deep Learning Approaches Reported in the Literature

Study	Base Models	Fusion Strategy	Dataset	Task	Reported Outcome	Major Limitation
Rahimzadeh & Attar, 2020 [13]	Xception, ResNet50V2	Feature-level concatenation	Cohen collection + RSNA pneumonia	Three-class	Reported class-wise results showing COVID-19 class performance below the aggregate figure	Aggregate metric conceals minority-class weakness; members not independently replaceable
Minaee et al., 2020 [12]	ResNet18, ResNet50, SqueezeNet, DenseNet121	Compared individually; sensitivity-specificity trade-off analysed	COVID-Xray-5k	Binary	Analysed the sensitivity-specificity trade-off explicitly	Small positive class; comparison rather than a combined model
Ismael & Sengur, 2021 [14]	ResNet18/50/101, VGG16/19 as feature extractors	Deep features passed to an SVM (hybrid, not multi-model fusion)	Small public CXR set	Binary	Found deep features competitive with end-to-end fine-tuning	Small dataset; classical head limits scalability
Chowdhury et al., 2020 [9]	Eight pretrained CNNs including DenseNet201, ResNet18	Compared individually, with and without augmentation	COVID-19 Radiography Database	Binary and three-class	Compared eight backbones and reported augmentation as beneficial	Curated single-source database; no cross-dataset evaluation
Punn & Agarwal, 2021 [15]	ResNet, Inception-ResNetV2, DenseNet169, NASNetLarge	Compared individually with oversampling and weighted loss	Public posteroanterior CXR	Three-class	Addressed class imbalance through oversampling and weighted loss	Limited posteroanterior images; no fusion of the compared models
Horry et al., 2020 [17]	VGG19 across three modalities	Per-modality transfer learning; modalities compared, not fused	X-ray, CT and ultrasound	Binary	Compared three modalities under a common protocol	Small COVID sample in each modality; no multimodal fusion

E. Comparative Performance

Tables 4 and 5 in Section VII summarise the reviewed studies. One point must be stated plainly before they are read: the accuracies reported in the primary literature are not directly comparable across studies, and treating them as a ranking would be a misreading of that literature. Section VII sets out the specific reasons.

VI. EXPLAINABLE AI FOR CHEST X-RAY CLASSIFICATION

A. Grad-CAM

Class activation mapping [25] showed that a network with global average pooling can localise discriminative regions without bounding-box supervision, but required a specific architecture. Grad-CAM [4] generalised the idea to arbitrary CNNs by weighting feature maps according to the gradient of the class score. It operates on the activations of the final convolutional layer, which retain spatial structure while encoding high-level semantics.

Let A^k denote the k-th feature map of that layer, with spatial indices (u, v), and let y^c denote the score for class c before the softmax. The importance weight of map k for class c is the gradient of the class score with respect to that map, averaged over its spatial extent:

$$\alpha_k{}^c = \frac{1}{Z}\sum_u \sum_v \frac{\partial y^c}{\partial A_{uv}{}^k} \tag{2}$$

where Z is the number of spatial locations. The localisation map is a weighted combination of the feature maps followed by a rectified linear unit:

$$L_{\text{Grad-CAM}}{}^c = \text{ReLU}(\sum_k \alpha_k{}^c A^k) \tag{3}$$

The rectifier is functionally necessary rather than cosmetic. Without it the map would include regions whose activation argues against class c alongside those that argue for it, and the two would be indistinguishable in the visualisation. The resulting map has the resolution of the final convolutional layer and is upsampled to the input size before being overlaid as a heatmap.

The limitations reported in the literature deserve emphasis. Grad-CAM maps are coarse, because the final convolutional layer has been heavily downsampled, so fine structure is not resolved. They are sensitive to the layer chosen: maps produced at different depths of the same network can differ substantially. Adebayo et al. [37] showed that some widely used saliency methods produce visually plausible maps that are largely insensitive to randomisation of the model parameters or the training labels, which means visual plausibility is not by itself evidence that a map reflects what the model learned. Accordingly, the correct statement is that Grad-CAM provides a visual indication of the image regions contributing to a model prediction. It does not demonstrate that the prediction is clinically correct, does not constitute a diagnosis, and cannot substitute for radiological assessment.

B. Other Explainability Approaches

Grad-CAM++ [38] extends Grad-CAM with a weighted combination of positive partial derivatives, which improves localisation where several instances of a class appear in one image and generally yields more complete object coverage. LIME [39] is model-agnostic: it perturbs the input, in images by toggling superpixels, and fits a sparse linear surrogate locally around the instance, so it can be applied to any classifier but produces explanations that depend on the perturbation and segmentation scheme and can be unstable between runs. SHAP [40] assigns each feature a Shapley value from cooperative game theory, giving attributions with desirable theoretical properties at substantially higher computational cost, and requiring approximation for deep networks. Integrated Gradients [41] attributes the prediction by integrating gradients along a path from a baseline input to the actual input, satisfying sensitivity and implementation-invariance axioms but requiring a baseline whose choice is not obvious for radiographs, where a zero image is not a natural neutral reference.

Within the COVID-19 radiograph literature specifically, Grad-CAM dominates by a wide margin, and comparative evaluation of attribution methods on this task is scarce.

Table 3. Explainable AI Methods Used in Medical Chest X-Ray Analysis

Method	Purpose	Advantages	Limitations	Application
Grad-CAM [4]	Class-discriminative localisation from final-layer gradients	Architecture-agnostic; single backward pass; widely implemented	Coarse resolution; sensitive to layer choice; plausibility is not evidence of correctness [37]	Dominant method in COVID-19 chest radiograph studies [5], [12]
Grad-CAM++ [38]	Refined localisation using positive partial derivatives	Better coverage of extended or multiple regions than Grad-CAM	Retains the coarse resolution and layer sensitivity of Grad-CAM	Radiograph and general medical image localisation
CAM [25]	Localisation via global average pooling weights	Conceptually simple; the basis for later gradient methods	Requires a specific architecture with global average pooling	Weakly supervised localisation in thoracic imaging
LIME [39]	Local surrogate model fitted to perturbed inputs	Model-agnostic; applicable to any classifier	Depends on segmentation and perturbation scheme; explanations can be unstable between runs	Instance-level explanation of radiograph classifiers
SHAP [40]	Shapley-value feature attribution	Theoretically grounded attribution with consistency properties	Computationally expensive; requires approximation for deep networks	Feature attribution in medical decision-support models
Integrated Gradients [41]	Path integral of gradients from a baseline to the input	Satisfies sensitivity and implementation-invariance axioms	Requires a baseline whose choice is not obvious for radiographs	Pixel-level attribution in medical image classification

C. Importance of Interpretability in Medical AI

Interpretability serves two distinct purposes in this domain, and the literature does not always separate them. The first is clinical: a reader deciding how much weight to give an automated output benefits from seeing what drove it. The second, and in the reviewed literature the more consequential, is diagnostic of the model rather than of the patient. A saliency map concentrated on image borders, text annotations, laterality markers or support hardware is direct evidence that the model is exploiting acquisition artefacts. DeGrave et al. [31] used exactly this kind of analysis to show that models trained on public COVID-19 radiograph datasets rely substantially on such shortcuts. Explainability is therefore best understood in this literature as an auditing tool for detecting dataset problems, and not as a validation of clinical correctness.

VII. COMPARATIVE ANALYSIS OF EXISTING STUDIES

The reviewed studies are summarised across two tables. Table 4 records the characteristics of each study, namely the dataset used, the task formulation, the architecture adopted and whether transfer learning was applied. Table 5 records the reported outcome and the principal methodological limitation of each. The separation is deliberate: study characteristics are the properties that determine whether two results are comparable at all, and presenting them alongside the outcomes invites exactly the direct numerical comparison that this section argues against.

A note on the performance columns of Table 5 is necessary. This review reproduces no numerical performance value from the primary sources. Metrics are marked NR throughout, indicating that the figure is not reported in this review and should be obtained from the cited publication directly. This is a deliberate editorial decision rather than an omission. Reproducing accuracy figures side by side in a single column invites readers to rank the studies, and the remainder of this section sets out why such a ranking would be unsound. Where a study's principal finding is qualitative and unambiguous, it is stated in words.

Table 4. Characteristics of the Reviewed Studies on COVID-19 Detection from Chest Radiographs

Author(s)	Year	Dataset Used	Task Formulation	Method / Architecture	Transfer Learning
Wang et al. [5]	2020	COVIDx (aggregated open sources)	Three-class: COVID-19, pneumonia, normal	COVID-Net, purpose-designed CNN	No
Ozturk et al. [7]	2020	Cohen collection and ChestX-ray8	Binary and three-class	DarkCovidNet, based on DarkNet-19	No
Apostolopoulos and Mpesiana [8]	2020	Composite public collection	Binary and three-class	VGG19, MobileNetV2 and other backbones	Yes
Narin et al. [6]	2021	Cohen collection and public Kaggle CXR	Binary, across four dataset configurations	ResNet50, InceptionV3, Inception-ResNetV2	Yes
Chowdhury et al. [9]	2020	COVID-19 Radiography Database	Binary and three-class	Eight pretrained CNNs, with and without augmentation	Yes
Rahman et al. [10]	2021	COVID-19 Radiography Database, extended	Binary and three-class	CNN backbones under varied image enhancement	Yes
Minaee et al. [12]	2020	COVID-Xray-5k	Binary	Deep-COVID: ResNet18, ResNet50, SqueezeNet, DenseNet121	Yes
Rahimzadeh and Attar [13]	2020	Cohen collection and RSNA pneumonia	Three-class	Feature-level concatenation of Xception and ResNet50V2	Yes
Ismael and Sengur [14]	2021	Small public CXR collection	Binary	Deep features from ResNet50 classified by SVM	Yes
Punn and Agarwal [15]	2021	Public posteroanterior CXR	Three-class	Fine-tuned CNNs with oversampling and weighted loss	Yes
Nishio et al. [16]	2020	Composite public collection	Three-class	VGG16 with combined augmentation strategies	Yes
Horry et al. [17]	2020	X-ray, CT and ultrasound	Binary, evaluated per modality	VGG19 transfer learning across three modalities	Yes
Tartaglione et al. [18]	2020	Multiple small public sources	Binary, cross-dataset evaluation	ResNet-based cross-dataset study	Yes
Maguolo and Nanni [20]	2021	Multiple public COVID-19 CXR sets	Binary, protocol critique	Standard backbones evaluated with lung fields masked out	Yes
DeGrave et al. [31]	2021	Public COVID-19 CXR datasets	Binary, with external evaluation	Saliency and generative analysis of learned features	Yes

Author(s)	Year	Dataset Used	Task Formulation	Method / Architecture	Transfer Learning
Roberts et al. [19]	2021	Systematic review of published models	Not applicable	Systematic review and risk-of-bias appraisal	Not applicable

Table 5. Reported Outcomes and Principal Limitations of the Reviewed Studies

Author(s)	Accuracy	Sensitivity	Specificity	F1-score	Principal Reported Outcome	Major Limitation
Wang et al. [5]	NR	NR	NR	NR	Introduced a purpose-designed architecture together with an open aggregated dataset and released both publicly	Aggregated multi-source data; class membership confounded with acquisition source
Ozturk et al. [7]	NR	NR	NR	NR	Reported the binary task as substantially easier than the three-class task; code released	Very small COVID-19 class; no external validation
Apostolopoulos and Mpesiana [8]	NR	NR	NR	NR	Found older and lightweight backbones competitive with deeper ones on a small composite dataset	Small positive class; single composite source; no cross-dataset test
Narin et al. [6]	NR	NR	NR	NR	Compared three backbones across four binary configurations without combining them	Severe class imbalance; results reported per configuration
Chowdhury et al. [9]	NR	NR	NR	NR	Compared eight backbones and reported augmentation as beneficial on a curated database	Single curated source; transferability to other institutions untested
Rahman et al. [10]	NR	NR	NR	NR	Demonstrated that the choice of image enhancement measurably changes classification outcomes	Implies that between-study variation is partly preprocessing rather than architecture
Minaee et al. [12]	NR	NR	NR	NR	Analysed the sensitivity-specificity trade-off explicitly rather than reporting accuracy alone	Few hundred COVID-19 images; specificity weaker than sensitivity
Rahimzadeh and Attar [13]	NR	NR	NR	NR	Reported class-wise results showing COVID-19 class performance well below the aggregate figure	Aggregate metric conceals minority-class weakness; members not independently replaceable
Ismael and Sengur [14]	NR	NR	NR	NR	Found deep features with a classical classifier competitive with end-to-end fine-tuning	Small dataset; classical head limits scalability

Author(s)	Accuracy	Sensitivity	Specificity	F1-score	Principal Reported Outcome	Major Limitation
Punn and Agarwal [15]	NR	NR	NR	NR	Addressed class imbalance directly through oversampling and weighted loss	Limited posteroanterior images; compared models not fused
Nishio et al. [16]	NR	NR	NR	NR	Showed that the combination of augmentation operations affects the result	Outcome depends strongly on augmentation choice, complicating comparison
Horry et al. [17]	NR	NR	NR	NR	Compared three imaging modalities under a common transfer-learning protocol	Small COVID-19 sample in each modality; modalities not fused
Tartaglione et al. [18]	NR	NR	NR	NR	Reported marked performance degradation under cross-dataset evaluation	Demonstrates poor generalisation across acquisition sources
Maguolo and Nanni [20]	NR	NR	NR	NR	Reported above-chance separation of the classes with the lung fields masked out entirely	Direct evidence of shortcut learning from source artefacts
DeGrave et al. [31]	NR	NR	NR	NR	Reported that models select acquisition shortcuts over radiographic signal, with performance falling on external data	Confirms that internal test accuracy is not evidence of clinical generalisation
Roberts et al. [19]	NR	NR	NR	NR	Concluded that none of the several hundred models reviewed was of potential clinical use	Identifies pervasive risk of bias, poor documentation and absent external validation

NR indicates that the metric is not reproduced in this review; readers should consult the cited publication for the reported figure. Studies differ in dataset, class definition, partitioning and evaluation protocol, and their results are not directly comparable.

A. Why Reported Performance Is Not Directly Comparable

Several factors, varying simultaneously across the reviewed studies, prevent published accuracies from being read as a ranking of methods.

- Dataset composition. Studies draw on different collections, and even those citing the same collection frequently use different versions with different numbers of COVID-19 images.
- Class definitions. Binary formulations separating COVID-19 from normal are systematically easier than three-class formulations that must also distinguish COVID-19 from other pneumonia, which is the clinically harder and more relevant distinction.
- Partitioning. Splits differ in proportion and, more importantly, in whether they are made at patient or image level. Image-level splitting places radiographs of the same patient in both training and test partitions and inflates reported metrics.
- Preprocessing. Rahman et al. [10] and Nishio et al. [16] both demonstrated that enhancement and augmentation choices measurably change results, so part of the between-study variation is preprocessing rather than architecture.
- Evaluation protocol. Studies differ in whether metrics are aggregate or per-class, whether cross-validation is used, and whether results are averaged over several random seeds. Single-split results without seed variation give no indication of measurement uncertainty.

- Internal versus external validation. Almost all reported metrics come from a held-out partition of the same corpus used for development. Tartaglione et al. [18] and DeGrave et al. [31] both showed that performance falls substantially under cross-dataset evaluation.

A further and more troubling observation is that differences of one or two percentage points between studies are frequently smaller than the variation attributable to partitioning and initialisation alone, particularly given positive classes numbering in the hundreds. Distinctions of this magnitude between architectures should not be treated as meaningful.

B. Critical Synthesis Across the Reviewed Literature

Reading across the studies rather than down the tables, several patterns emerge that individual papers do not reveal in isolation.

1) Dataset differences dominate architectural differences

Reported performance tracks the dataset and the task formulation far more closely than it tracks the network. Studies using curated single-source collections report higher figures than those using multi-source composites, and those evaluating across datasets report lower figures still [18], [31], irrespective of backbone. Because no two studies in the reviewed set share a dataset, a partition and a preprocessing pipeline simultaneously, the effect of architecture cannot be isolated from published results. The reviewed literature therefore does not establish that any one backbone is better suited to this task than another, and claims to that effect in individual papers rest on comparisons confounded by data.

2) Binary and multiclass formulations are not equivalent tasks

Studies reporting both formulations consistently report the binary task as easier than the three-class task [7], [8], [9]. This is expected, since distinguishing COVID-19 pneumonia from other pneumonia is a finer discrimination than distinguishing it from normal anatomy, and it is also the clinically meaningful one: a patient presenting with radiographic pneumonia needs the aetiology resolved, not the presence of pneumonia confirmed. Binary COVID-versus-normal figures therefore overstate what the same model would achieve in the setting where it would actually be deployed, and comparing a binary figure from one study against a three-class figure from another is not meaningful.

3) Class imbalance and the inadequacy of aggregate metrics

The COVID-19 class is a small minority in every reviewed corpus, so aggregate accuracy is a weak summary. Where per-class metrics are reported, minority-class performance is consistently weaker than the aggregate figure, most visibly in the class-wise results of Rahimzadeh and Attar [13]. Studies reporting only aggregate accuracy therefore cannot be assessed on the metric that matters most clinically, namely sensitivity to the positive class. Mitigations reported in the literature, including oversampling and weighted loss [15], address the training-side effect but do not make aggregate accuracy an informative metric.

4) Patient-level versus image-level partitioning

Where a corpus contains several radiographs of the same patient, splitting at image level places correlated images in both training and test partitions, allowing the model to recognise the patient rather than the pathology and biasing the reported metric upwards. Many reviewed studies do not state which level of separation was used, so the presence or absence of this leakage cannot be determined from the publication. This is a reporting failure with direct consequences for interpretation: a result obtained under image-level splitting is not comparable with one obtained under patient-level splitting, and readers frequently cannot tell which they are looking at.

5) Preprocessing as an uncontrolled variable

Two studies varied preprocessing deliberately and both found that it changes the outcome: Rahman et al. [10] for image enhancement and Nishio et al. [16] for augmentation strategy. It follows that part of the between-study variation attributed in the literature to architecture is in fact preprocessing. Because most studies neither vary preprocessing systematically nor describe it in reproducible detail, this contribution cannot be quantified from published work, and it constitutes an uncontrolled variable running through the entire comparative literature.

6) Transfer learning is universal and therefore untested

Almost every reviewed study initialises from ImageNet weights [22], which means the literature contains few controlled comparisons against training from scratch on the same data. Transfer learning is thus a shared assumption of the field rather than a hypothesis it has tested. Given the substantial domain shift between natural photographs and grayscale radiographs, and given the very small target datasets, the benefit may

derive as much from regularisation against overfitting as from transferable pathological features, and the reviewed studies do not distinguish between these explanations.

7) Ensemble gains are plausible but weakly evidenced

Ensemble and hybrid methods are generally reported as improving on their individual members, which is consistent with established theory [23], [24], [29]. The evidential basis is nonetheless weaker than it appears. Improvements are typically reported from a single partition without variation across random seeds, so it cannot be determined whether a reported margin exceeds the measurement noise. With positive classes numbering in the hundreds, seed-to-seed and split-to-split variation can plausibly account for differences of several percentage points. Reported ensemble margins of that order should therefore be treated as suggestive rather than demonstrated.

8) Internal validation, dataset-source bias and shortcut learning

These three issues compound one another and together constitute the most serious weakness in the reviewed literature. Almost all reported performance is internal, obtained on a held-out partition of the development corpus. Many of those corpora are composites in which the positive and negative classes originate from different institutions and equipment, so class membership is confounded with acquisition source. A network minimising classification loss has no reason to prefer pathological features over source-specific ones, and the evidence that it does not is direct: Maguolo and Nanni [20] obtained above-chance separation with the lung fields masked out entirely, and DeGrave et al. [31] showed that models select acquisition shortcuts over radiographic signal, with performance falling on external data. Tartaglione et al. [18] documented the same degradation under cross-dataset evaluation. A high internal accuracy is therefore consistent both with a model that has learned pathology and with one that has learned to recognise its data source, and the internal result cannot distinguish the two.

9) Explainability is used illustratively rather than analytically

Grad-CAM appears in most reviewed studies, but typically as a small panel of favourable examples rather than as a systematic analysis. The known limitations of saliency methods, including coarse resolution, sensitivity to layer choice, and the finding of Adebayo et al. [37] that plausible-looking maps can be largely insensitive to model randomisation, are rarely acknowledged. The literature that used attribution analytically rather than illustratively [31] is precisely the literature that uncovered the shortcut problem, which suggests the method is more valuable as a model-auditing tool than as the reassurance it is usually deployed as.

10) Reproducibility and clinical validation

Reproducibility varies widely. A minority of studies release code, weights and data [5], [7]; many report neither random seeds nor library versions nor the dataset version used, and since the principal public datasets were revised repeatedly during the period in question, a citation by name alone is often insufficient to identify what was used. Beyond reproducibility lies the larger gap: essentially none of the reviewed work has been evaluated prospectively in clinical practice, compared against reporting radiologists in situ, or assessed for effect on patient outcomes. Roberts et al. [19] reached this conclusion across several hundred studies. The distance between reported classification accuracy and demonstrated clinical utility remains the defining limitation of this field.

VIII. CHALLENGES AND RESEARCH GAPS

A. Dataset Size and Class Imbalance

The COVID-19 positive class in the public collections is small in absolute terms and small relative to the negative classes. Deep networks with millions of parameters trained on positive classes of a few hundred images are prone to overfitting, and imbalance biases the classifier towards the majority class while inflating aggregate accuracy. Mitigations reported in the literature reduce but do not remove the effect.

B. Dataset-Source Bias and Shortcut Learning

This is the most serious methodological problem identified in the reviewed literature and warrants detailed treatment. Several widely used COVID-19 radiograph datasets are composites in which the positive and negative classes originate from different sources. The COVID-19 images were frequently collected from published case reports and online repositories, often already compressed, rescaled or annotated, while the normal and pneumonia images came from pre-existing hospital datasets acquired on different equipment under different protocols. Class membership is therefore confounded with acquisition source.

A network minimising classification loss has no reason to prefer pathological features over source-specific ones if the latter separate the classes more easily, and the evidence that this occurs is direct rather than

speculative. Maguolo and Nanni [20] showed that classifiers achieve above-chance separation of COVID-19 from non-COVID-19 radiographs even when the lung fields are masked out entirely, which means the discriminative signal lies outside the anatomy of interest. DeGrave et al. [31] reached a consistent conclusion using saliency and generative analysis, showing that models rely on laterality markers, image borders, text annotation and other acquisition artefacts, and that performance falls sharply on external data.

The consequence for reading this literature is significant. A high accuracy on a held-out partition of a composite public dataset is consistent both with a model that has learned radiographic pathology and with one that has learned to recognise the data source, and the internal test result cannot distinguish between them. Only external validation on data acquired independently, or deliberate ablation such as lung-field masking, can separate the two. Reported accuracy on public composites should therefore be treated as an upper bound obtained under favourable conditions rather than as an estimate of clinical performance.

C. Data Leakage

Where a dataset contains multiple radiographs of the same patient, splitting at image level places correlated images in both training and test partitions. The model can then recognise the patient rather than the pathology, and the reported metric is optimistically biased. Many reviewed studies do not state whether their splits were made at patient or image level, which makes the presence or absence of this problem unverifiable from the publication alone.

D. Lack of External Validation

Almost all reported performance is internal. Roberts et al. [19] systematically reviewed several hundred papers on machine learning for COVID-19 imaging and concluded that none of the models reviewed was of potential clinical use, citing small samples, poor documentation, high risk of bias and a near-total absence of external validation. Tartaglione et al. [18] demonstrated the resulting degradation directly.

E. Heterogeneous Evaluation Protocols and Absent Benchmarks

Studies differ in metrics reported, in aggregate versus per-class reporting, in cross-validation, and in whether seed variation is assessed. There is no standard benchmark with a fixed, patient-level, publicly published split for this task, which is precisely what would make results comparable. Its absence is a structural gap in the field rather than a shortcoming of any individual study.

F. Reproducibility

Reproducibility varies substantially. Some studies release code, data and weights [5], [7]; many report neither random seeds nor library versions nor the exact dataset version used. Given that dataset versions changed during the period in question, a citation by name alone is often insufficient to identify what was used.

G. Interpretability, Computational Cost and Clinical Validation

Explanation is frequently presented as a post-hoc illustration on a few selected images rather than as a systematic analysis, and the limitations of saliency methods [37] are rarely acknowledged. Ensemble and hybrid methods multiply training and inference cost, which is in tension with the resource-limited settings that motivate radiograph-based screening. Above all, essentially none of the reviewed work has been evaluated prospectively in a clinical setting, compared against reporting radiologists in practice, or assessed for effect on patient outcomes. The gap between reported classification accuracy and demonstrated clinical utility remains the largest in the field.

IX. FUTURE DIRECTIONS

The directions below follow from the gaps identified in Section VIII and are drawn from the recommendations made in the reviewed literature. None is presented as completed work.

1) Larger and better-balanced datasets, with the positive class large enough that minority-class sensitivity can be estimated with reasonable precision.

2) Multi-centre datasets in which each class is drawn from the same distribution of sources, so that class membership is not confounded with acquisition source.

3) Routine external validation on independent data, with performance reported separately by acquisition site, as recommended by Roberts et al. [19].

4) Standardised public benchmarks with fixed, patient-level, published splits, so that reported results become comparable across studies.

5) Federated learning, which allows models to be trained across institutions without centralising patient data and thereby addresses both the scale and the privacy constraints simultaneously [42].

6) Lightweight architectures and knowledge distillation of ensembles into single compact models, addressing the computational cost of multi-model approaches.

7) Attention mechanisms, evaluated not only for accuracy but for whether they improve the anatomical plausibility of the resulting explanations.

8) Multimodal imaging that combines radiography with computed tomography or ultrasound, extending the comparison reported by Horry et al. [17].

9) Systematic comparison of explainability methods on this task, including the sanity checks of Adebayo et al. [37], rather than the current reliance on Grad-CAM without validation.

10) Model calibration and uncertainty estimation, since a screening system whose confidence estimates are unreliable cannot be thresholded appropriately; modern networks are known to be poorly calibrated by default [43].

11) Deliberate bias auditing, including lung-field masking and cross-source ablation of the kind used in [20] and [31], as a routine reporting requirement rather than an occasional critique.

12) Prospective clinical evaluation, including comparison against reporting radiologists and assessment of the effect on triage decisions, before deployment is contemplated.

X. CONCLUSION

This paper has reviewed the literature on deep learning for COVID-19 detection from chest X-ray images. Deep learning, and CNNs in particular, has been investigated extensively for this task, and transfer learning from ImageNet-pretrained backbones is the dominant approach, adopted in response to the small size of the available datasets. VGG, ResNet and DenseNet architectures recur most frequently, with lightweight and multi-scale families investigated for constrained deployment. Ensemble and hybrid methods have been proposed to provide architectural diversity and reduce dependence on a single model, and are generally reported as improving on their individual members, consistent with established ensemble theory. Explainability, chiefly through Grad-CAM, has become an expected component, although its reported use is often illustrative rather than systematic.

The central finding of this review is that published performance varies substantially with dataset, task formulation, partitioning, preprocessing and evaluation protocol, and that reported accuracies are consequently not directly comparable across studies. Differences of a few percentage points between architectures are frequently smaller than the variation attributable to methodological choices, and should not be read as evidence of architectural superiority. More seriously, dataset construction in this field has often confounded class membership with acquisition source, and direct evidence exists that models exploit that confound rather than radiographic pathology [20], [31]. Combined with the near-absence of external validation documented by Roberts et al. [19], this means that high internal accuracy on public composites should be interpreted as an upper bound obtained under favourable conditions rather than as an estimate of clinical performance.

The techniques themselves are not thereby invalidated. What the literature demonstrates is that the principal constraints on progress are now the data and the evaluation protocol rather than the network architecture. Larger, properly constructed multi-centre datasets, standardised patient-level benchmarks, routine external validation, deliberate bias auditing, calibrated and uncertainty-aware outputs, and prospective clinical evaluation are the developments most likely to move the field from high reported accuracy towards demonstrated clinical utility.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[2] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 4700-4708.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. 3rd Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, 2015.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626.

[5] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," Scientific Reports, vol. 10, no. 1, art. 19549, 2020.

[6] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," Pattern Analysis and Applications, vol. 24, no. 3, pp. 1207-1220, 2021.

[7] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," Computers in Biology and Medicine, vol. 121, art. 103792, 2020.

[8] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," Physical and Engineering Sciences in Medicine, vol. 43, no. 2, pp. 635-640, 2020.

[9] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi, M. B. I. Reaz, and M. T. Islam, "Can AI help in screening viral and COVID-19 pneumonia?," IEEE Access, vol. 8, pp. 132665-132676, 2020.

[10] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. H. Chowdhury, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," Computers in Biology and Medicine, vol. 132, art. 104319, 2021.

[11] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "COVID-19 image data collection: Prospective predictions are the future," arXiv preprint arXiv:2006.11988, 2020.

[12] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," Medical Image Analysis, vol. 65, art. 101794, 2020.

[13] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2," Informatics in Medicine Unlocked, vol. 19, art. 100360, 2020.

[14] A. M. Ismael and A. Sengur, "Deep learning approaches for COVID-19 detection based on chest X-ray images," Expert Systems with Applications, vol. 164, art. 114054, 2021.

[15] N. S. Punn and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks," Applied Intelligence, vol. 51, no. 5, pp. 2689-2702, 2021.

[16] M. Nishio, S. Noguchi, H. Matsuo, and T. Murakami, "Automatic classification between COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy on chest X-ray image: combination of data augmentation methods," Scientific Reports, vol. 10, art. 17532, 2020.

[17] M. J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, and N. Shukla, "COVID-19 detection through transfer learning using multimodal imaging data," IEEE Access, vol. 8, pp. 149808-149824, 2020.

[18] E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, and M. Grangetto, "Unveiling COVID-19 from chest X-ray with deep learning: A hurdles race with small data," International Journal of Environmental Research and Public Health, vol. 17, no. 18, art. 6933, 2020.

[19] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. H. F. Rudd, E. Sala, and C.-B. Schonlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, vol. 3, no. 3, pp. 199-217, 2021.

[20] G. Maguolo and L. Nanni, "A critic evaluation of methods for COVID-19 automatic detection from X-ray images," Information Fusion, vol. 76, pp. 1-7, 2021.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 2009, pp. 248-255.

[22] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.

[23] L. K. Hansen and P. Salamon, "Neural network ensembles," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993-1001, 1990.

[24] T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems (MCS 2000), Lecture Notes in Computer Science, vol. 1857, Berlin, Germany: Springer, 2000, pp. 1-15.

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2921-2929.

[26] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics Gems IV, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474-485.

[27] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2097-2106.

[28] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[29] R. Polikar, "Ensemble based systems in decision making," IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21-45, 2006.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818-2826.

[31] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," Nature Machine Intelligence, vol. 3, no. 7, pp. 610-619, 2021.

[32] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1251-1258.

[34] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), Long Beach, CA, USA, 2019, pp. 6105-6114.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9.

[36] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artificial Intelligence, Honolulu, HI, USA, 2019, pp. 590-597.

[37] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in Advances in Neural Information Processing Systems (NeurIPS), Montreal, Canada, 2018, pp. 9505-9515.

[38] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 2018, pp. 839-847.

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135-1144.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 4765-4774.

[41] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. 34th Int. Conf. Machine Learning (ICML), Sydney, Australia, 2017, pp. 3319-3328.

[42] G. A. Kaissis, M. R. Makowski, D. Rueckert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nature Machine Intelligence, vol. 2, no. 6, pp. 305-311, 2020.

[43] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning (ICML), Sydney, Australia, 2017, pp. 1321-1330.