Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep Learning for Lesion Classification in CT Imaging: A Systematic Literature Review

UMMEY HANY AINAN
Universiti Malaya
NOR ANIZA ABDULLAH
Universiti Malaya
K. A.S.H. KULATHILAKE
kule@as.rjt.ac.lk

Rajarata University of Sri Lanka
AZNUL QALID Md SABRI
Universiti Malaya
JEANNIE WONG
Universiti Malaya
KHIN WEE LAI
Universiti Malaya
SUZAN J. OBAIYS
Universiti Malaya

Systematic Review

Keywords: Deep learning, lesion classification, computed tomography imaging, low-dose computed tomography, convolutional neural network, transformer

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10632411/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Accurate lesion classification in computed tomography (CT) imaging is crucial for early diagnosis and effective clinical decision-making. With the growing adoption of low-dose CT (LDCT) to reduce radiation exposure and the rapid evolution of deep learning (DL), automated lesion analysis has gained significant momentum. Despite this progress, existing research often lacks a unified assessment of model design, generalizability, and training practices. This systematic literature review (SLR) analyzes recent DL approaches for lesion classification from CT images, focusing on the influence of model design and training strategies on performance and generalizability. A comprehensive search was conducted across Web of Science, PubMed, Scopus, and IEEE Xplore. Following a rigorous screening and selection process in accordance with PRISMA guidelines, 57 peer-reviewed articles published between January 2020 and December 2025 were included. The findings show that lung and abdominal lesions were the primary research focus, with CNN-based models remaining the dominant approach because of their effectiveness in learning local morphological and textural features. Transformer-based models demonstrated superior contextual learning, while hybrid architectures generally achieved the best performance by integrating complementary local and global feature representations. Self-supervised and transfer learning further improved generalizability in data-limited settings. However, most studies relied on retrospective, small, and class-imbalanced datasets with limited external validation, restricting reproducibility and clinical translation. Persistent challenges include the limited availability of public LDCT datasets, insufficient model interpretability, and inadequate evaluation under real-world clinical conditions. Future research should prioritize large multi-center datasets, robust external validation, explainable AI, and clinically representative evaluation frameworks to develop reliable and generalizable CT-based lesion classification systems.

1 Introduction

According to the world health organization (WHO), cancer remains one of the leading causes of death globally, responsible for nearly one in six deaths, with an estimated 20 million new cases and 9.7 million deaths reported in 2022 [1] The global cancer burden is projected to increase substantially, with incidence expected to rise by approximately 77% by 2050, reaching an estimated 35 million new cases [1]. This alarming trend underscores the urgent need for improved strategies in cancer diagnosis, prevention, and treatment. Among available imaging modalities, computed tomography (CT) serves as a crucial diagnostic tool, providing high-resolution cross-sectional anatomical images that facilitate early lesion detection, accurate tumor staging, and effective treatment monitoring [2]. However, despite its diagnostic advantages, CT imaging involves exposure to ionizing radiation [2]. In clinical practice, low-dose CT (LDCT) is often preferred to minimize patient exposure to ionizing radiation while maintaining diagnostic image quality.

In medical imaging, lesion classification refers to the process of distinguishing between benign and malignant abnormalities and characterizing lesion types based on shape, texture, and enhancement patterns visible in CT [3]. Accurate lesion classification is fundamental for early cancer diagnosis, treatment planning, and prognostic evaluation. However, manual classification by radiologists can be subjective and time-consuming, leading to inter-observer variability. To address these challenges, computer-aided diagnosis (CAD) systems have been developed to assist radiologists in classifying lesions with greater accuracy and consistency by analyzing quantitative imaging features such as radiomic patterns and texture descriptors [2].

Early CAD systems relied on handcrafted features and rule-based classifiers, whose performance was constrained by the difficulty of manually specifying features that generalize across patients, scanners, and acquisition protocols [4]. The advent of deep learning (DL) shifted this paradigm: rather than depending on predefined descriptors, DL models learn hierarchical representations directly from image data, capturing both fine local detail and broader contextual patterns relevant to lesion differentiation [5–7]. These properties are especially advantageous in LDCT, where noise and quality degradation obscure fine structure; DL models can exploit contextual cues from surrounding tissue to recognize irregular margins and heterogeneous textures that may be imperceptible to human observers [8, 9]. As a result, DL has become the dominant approach for CT-based lesion classification and a cornerstone of data-driven, scalable precision oncology.

Despite rapid advances in deep learning for medical imaging, a critical review of existing surveys reveals that very few recent reviews comprehensively address lesion classification from CT images across multiple organs [10]. Most existing reviews are narrowly focused on specific lesion types (e.g., lung, liver lesions) [11–14]. Moreover, although deep learning–based approaches are widely reported, current reviews often lack a systematic discussion of model architectures, training strategies, generalization across diverse datasets, and the clinical relevance of commonly reported evaluation metrics, particularly from a multi-organ perspective [15]. Motivated by the identified gap, this systematic literature review critically examines DL-based lesion classification from CT image studies published between 2020 and 2025. The review draws from several reputable academic databases known for their scholarly articles and studies. It is guided by the following research questions (RQ):

- RQ1: Which DL models are most commonly applied for lesion classification in CT scans of lung and abdominal organs?
- RQ2: How do choices in model architecture influence classification accuracy and generalization across various datasets?
- RQ3: What training strategies are commonly employed to improve model performance?
- RQ4: What are the known challenges and limitations of the quantitative metrics commonly reported in studies on CT-based lesion classification using DL?

Additionally, this review adopts a multi-organ perspective to provide a more comprehensive understanding of how DL methods are applied across various anatomical contexts. The analysis primarily focuses on lung and abdominal organs, where CT remains the dominant imaging modality for lesion detection, screening, and diagnosis. In contrast, lesion classification for several other organs is more commonly investigated using alternative modalities, such as magnetic resonance imaging (MRI) for brain tumors [5], ultrasound and mammography for breast lesions [6], and dermoscopy or optical imaging for skin lesions [7]. Consequently, relatively few studies have explored CT-based lesion classification outside thoracic and abdominal regions. While many existing reviews focus narrowly on single-organ lesion classification, a multi-organ approach allows for the identification of shared challenges, transferable modeling

strategies, and diverse data characteristics. This scope is significant for guiding the development of generalizable and clinically relevant DL models. Hence, the key contributions of this review are as follows:

- It systematically identifies and categorizes the most frequently employed DL models for lesion classification across lung and abdominal organs using CT imaging. We also highlight organ-specific model preferences and general trends.
- It analyzes the influence of model architecture and training strategies (e.g., data augmentation, transfer learning) on classification performance and model generalizability across diverse datasets, and also discusses the implications of architectural complexity for the clinical feasibility of these models.
- It highlights the evaluation metrics (e.g., AUC, accuracy, sensitivity, F1-score) used in DL-based lesion classification studies and discusses their limitations and applicability in clinical practice.
- It synthesizes common challenges and methodological gaps in the current literature and suggests future research directions to improve DL-based lesion classification in CT imaging.

The review is organized as follows: The research method is presented in the subsequent section. The results of the systematic literature review following PRISMA guidelines are presented in Section 3. The discussion of the literature review is reported in Section 4. Finally, Section 5 concludes the review.

2 Methods

This section describes the methodology used to conduct a systematic literature review on DL-based lesion classification across multiple organs. The review process followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [16], ensuring a systematic, transparent, and replicable approach.

2.1 Identifying information sources

To ensure the inclusion of high-quality and relevant literature, a systematic search was conducted in the Web of Science, PubMed, Scopus, and IEEE Xplore databases. These reputable academic sources were selected for their extensive coverage of peer-reviewed research. The search strategy was carefully designed to identify journal articles that align with the objectives of this review, ensuring a comprehensive and credible evidence base.

2.2 Search strategy

The search was restricted to studies published between January 2020 and December 2025, reflecting the period of rapid advancement in CNNs and related DL techniques for CT imaging. A structured search strategy was designed in line with the research questions, combining keywords related to lesion classification, deep learning methods, CT/LDCT imaging, and multiple organs. Boolean operators (AND/OR) were used to refine the search. The core query included:

(“lesion classification” OR “nodule classification” OR “lesion diagnosis”) AND (“deep learning” OR “CNN” OR “transfer learning” OR “pre-trained model” OR “hybrid model” OR “Transformer” OR “vision Transformer” OR “self-supervised learning model”) AND (“multi-organ” OR “lung” OR “abdomen” OR “liver” OR “pancreas” OR “stomach” OR “kidney”) AND (“CT image” OR “LDCT image”)

2.3 Eligibility criteria

The inclusion and exclusion criteria were systematically developed to determine which studies were eligible for inclusion in the review and which were to be excluded. Table 1 presents a detailed overview of the specific parameters applied for study selection.

Table 1
Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Studies published between January 2020 and December 2025	Studies not involving DL (e.g., purely classical machine learning).
Articles that apply DL techniques for lesion classification.	Studies lacking sufficient methodological details or performance evaluation metrics
Peer-reviewed papers, articles.	Articles not written in English.
	Studies using imaging modalities other than CT
	Review articles

2.4 Study selection process

The study selection was conducted through a structured four-phase process: identification, screening, eligibility, and inclusion, following the PRISMA framework. This multi-stage process ensured that only relevant, high-quality studies aligned with the research questions were retained for review. A visual summary of the process is presented in Fig. 1.

Identification: An initial search was conducted in the Web of Science, PubMed, Scopus, and IEEE Xplore databases using a predefined Boolean query based on lesion classification, deep learning, CT imaging, and multi-organ systems. This search yielded 3142 records. All retrieved records were exported to a shared Google Sheet, where bibliographic data (title, authors, DOI, publication year, and abstract) were documented for transparency and team collaboration.

Screening: After removing 840 duplicate records, the remaining studies were screened based on titles and abstracts. 2,164 records were excluded due to lack of relevance to deep learning–based lesion classification in CT images, leaving 138 studies for full-text assessment.

Eligibility: Two independent reviewers conducted full-text assessments using predefined inclusion and exclusion criteria. Studies were excluded at this stage due to reasons such as insufficient methodological or evaluation detail or review articles. A total of 81 studies were excluded.

Inclusion: The remaining 57 studies were reviewed and validated by the broader research team. Only those that met all inclusion criteria and aligned thematically with the review objectives were retained for qualitative synthesis.

2.5 Quality assessment

To ensure the methodological rigor of the included studies, a quality assessment was conducted using a modified version of the Downs and Black checklist [17], which is widely used for both randomized and non-randomized studies, including those involving artificial intelligence and medical imaging. Eight criteria were scored (0 = not addressed, 1 = partially, 2 = fully addressed), covering clarity of objectives, reported outcomes, dataset, method transparency, evaluation strategy, bias, generalizability, and reproducibility. Two independent reviewers assessed each study and resolved discrepancies through discussion. Studies were categorized as high (13–16), moderate (9–12), or low quality (< 9). Only studies rated as moderate or high quality were included. Among the included studies, 8% are moderate, and 92% are high.

2.6 Data extraction

Data extraction was conducted independently by two reviewers using a standardized form. Key information was collected to address the research questions and ensure consistency across studies. Extracted data included: bibliographic details, organ type, dataset characteristics, model architecture, training strategies, evaluation metrics, key findings, and reported limitations. Discrepancies were resolved through discussion or by a third reviewer. The extracted data supported structured analysis and comparison across the included studies.

3 Results

The initial search conducted in several databases identified 3142 records. After removing duplicates and irrelevant studies during the title and abstract screening phase, 138 full-text articles were assessed for eligibility. Following a detailed full-text review according to predefined inclusion and exclusion criteria, 81 articles were excluded due to factors such as insufficient methodological or evaluation detail or review articles. Ultimately, 57 studies were included in the final review, as illustrated in Fig. 1.

3.1 Lesion diversity and organ-specific focus in CT imaging

On CT scans, lesions show different characteristics depending on the organs, malignancy risk, and radiological features such as size, shape, margins, density, and contrast enhancement [3], [18], as shown in Table 2. Both radiologists and DL models depend on these lesion characteristics for accurate lesion detection. Accordingly, effective learning of these lesion characteristics is essential for DL models to achieve performance comparable to that of radiologists [3]. From the perspective of malignancy risk, lesions can be classified as benign (e.g., cysts, hemangiomas), premalignant (e.g., subsolid nodules, intraductal papillary mucinous neoplasm (IPMN), mucinous cystic neoplasm (MCN)), or malignant (e.g., pancreatic ductal adenocarcinoma (PDAC), hepatocellular carcinoma (HCC), Renal cell carcinoma (RCC), metastases). However, detecting these lesions is difficult due to low contrast, poorly defined edges, overlapping features, and so on, as shown in Table 3.

According to our analysis of the literature, cancers in the lung and abdominal regions emerge as critical focus areas in CT-based lesion analysis. Analysis of the reviewed studies shows that the lung (41%) and abdomen (28%) were the most frequently investigated anatomical sites for CT-based lesion classification using DL. This emphasis reflects their high disease prevalence, routine use of CT imaging for diagnosis and screening, and the complex lesion characteristics of abdominal organs (e.g., liver, stomach, pancreas, and kidneys), which make automated analysis both clinically important and technically challenging. Accordingly, this review places particular emphasis on these two regions to reflect the prevailing research trends and clinical importance in automated lesion detection.

Table 2
Normal vs lesion characteristics in CT imaging

Tissue / Lesion Type	Shape	Edges / Margins	Density (Before Contrast)	Enhancement (After Contrast)
Normal tissue	Smooth, natural	Clear, smooth	Same as organ	Brightens evenly
Lesion (general)	Round, oval, lobulated, or irregular	Benign: smooth, Malignant: uneven/blurred	Appears darker than the surrounding tissue (darker), appears whiter than the surrounding tissue (brighter), or some parts are dark, some are bright (mixed)	Benign: even or edge only (uniform enhancement throughout the lesion or enhancement limited to the outer rim.) Malignant: uneven, bright early, then fades, may have dead spots (areas that stay dark)
HCC [19]	Round or irregular	Uneven/blurred	Darker or the same as liver	Bright early, then fades (washout)
ICC	Irregular	Uneven/blurred	Darker	Edge bright first, fills slowly
Metastasis [19]	Round or variable	Uneven/blurred	Darker	Edge bright, inside uneven
Cyst [19]	Round, smooth	Smooth	Darker	No change (stays dark)
Hemangioma [19]	Round or lobulated	Smooth	Darker	Edge bright first, fills to the center
FNH [19]	Round or lobulated	Smooth	Darker or the same	Very bright early, fades evenly
Solid pulmonary nodule [20]	Round	Smooth	Darker or mixed	Uneven, may have dead spots
Part-solid nodule [21]	Round	Smooth to slightly irregular	Mixed	Even or edge only
ADC [22]	Irregular/lobulated	Uneven/blurred	Mixed	Minimal or peripheral enhancement
SCC [22]	Lobulated	Uneven/blurred	Mixed	Uneven, may have dead spots
CIP, PLC, RP [23]	Patchy, blurry spots	Blurred	Hazy or mixed	No appreciable enhancement
EGC [24]	Focal wall thickening	Smooth to mildly irregular	Same or slightly darker than the stomach wall	Mild to moderate, mostly in the inner layer
AGC [24]	Irregular thickening or mass	Irregular or blurred	Mixed or darker, may have necrotic areas	Uneven, may fade or show invasion into fat or nearby organs
IPMN [25]	Smooth	Smooth or lobulated	Darker	Minimal or no enhancement
MCN [25]	Round or oval	Smooth	Darker	Even or edge only
SCN [25]	Lobulated	Smooth	Darker	Septal and central scar enhancement
SPN [25]	Mixed solid-cystic	Smooth	Mixed	Uneven, may have dead spots
ccRCC [26]	Irregular or round	Blurry or lobulated	Mixed; may have darker areas	Strong early enhancement with washout
pRCC [26]	Oval or round	Smooth	Homogeneous, brighter	Mild and gradual enhancement
chRCC [26]	Round or lobulated	Smooth	Homogeneous	Moderate, delayed enhancement

Note: HCC= hepatocellular carcinoma, ICC= intrahepatic cholangiocarcinoma, FNH= focal nodular hyperplasia, ADC= adenocarcinoma, SCC= squamous cell carcinoma, CIP= cryptogenic organizing pneumonia, PLC= pulmonary lymphangitic carcinomatosis, RP= radiation pneumonitis, EGC= early gastric cancer, AGC= advanced gastric cancer, IPMN= intraductal papillary mucinous neoplasm, MCN= mucinous cystic neoplasm, SCN= serous cystic neoplasm, SPN= solid pseudopapillary neoplasm, ccRCC= clear cell renal cell carcinoma, pRCC= papillary renal cell carcinoma, chRCC= chromophobe renal cell carcinoma

Table 3
Detection challenges and model impacts by organ and lesion type

Organ	Lesion Type	Detection Challenge	Model Implication	Reference
Lung	Solid pulmonary nodule	Low contrast, slight lobulation.	Subtle features and unclear boundaries limit feature extraction and increase the risk of misclassification.	[20]
	CIP, PLC, RP	Overlapping patterns, diffuse boundaries, and low contrast.	Ambiguity reduces class separability and feature reliability.	[23]
	Part-solid nodule	Diffuse boundaries	Blurred margins hinder lesion localization.	[21]
	ADC, SCC	Overlapping patterns	Subtle differences increase misclassification risk.	[22]
Liver	HCC, ICC	Overlapping patterns, diffuse boundaries.	Low contrast or similarity causes boundary loss and feature confusion.	[19]
	Cyst	Low contrast		
	Hemangioma	Diffuse boundaries.		
	Metastasis	Small, overlapping patterns, low contrast		
	FNH	Overlapping patterns		
Stomach	EGC	Low contrast, overlapping patterns.	Minimal structural distortion may reduce the sensitivity of automated detection models.	[24]
	AGC	Diffuse boundaries.	Boundary ambiguity increases classification errors.	
Pancreas	IPMN, MCN, SCN, SPN.	Overlapping patterns.	Visual similarity lowers class specificity.	[25]
Kidney	ccRCC	Overlapping patterns	Feature loss or false positives due to similarity, low contrast, or calcifications.	[26]
	pRCC	Low contrast		
	chRCC	Diffuse boundaries		

3.2 Dataset characteristics in CT-based lesion classification

The reviewed studies use a diverse set of CT datasets covering lung, liver, kidney, and abdominal lesions. As shown in Tables 4 and 5, these datasets differ widely in size, organ focus, annotation detail, and class balance. Large public datasets such as LIDC-IDRI, LUNA16, and NLST enable broader model development, whereas many abdominal and multi-center datasets are private, reflecting real-world data-access limitations. Class imbalance is common across both public and private datasets, where malignant or rare lesion cases are often underrepresented compared to benign cases, and lesion-level annotations are not always consistently reported, making training and evaluation more difficult. In this review, balanced datasets contain a similar number of benign and malignant cases; nearly balanced datasets show only a slight difference between the two, imbalanced datasets are dominated by one class (usually benign), and highly imbalanced datasets contain very few malignant samples compared to benign ones. While few public datasets are nearly balanced, many are imbalanced or highly imbalanced, and most private datasets are imbalanced or do not clearly report their class distribution, as reflected by missing entries in Table 5. Notably, public access to LDCT datasets is limited, which restricts reproducibility and slows progress in low-dose CT research.

Table 4
Public dataset summary for CT-based lesion classification

Dataset ID	Dataset Name	Link	Modality	Centre(s)	Sample	Balanced/ Imbalanced	Lesion Level Annotation for classification	In Fo
D01	LIDC-IDRI [27]	https://www.cancerimagingarchive.net/collection/lidc-idri/	CT	Multi	1018	Highly imbalanced	Yes	DI
D02	LUNA16 [34]	https://luna16.grand-challenge.org/Data/	CT	Multi	888	Imbalanced	No	DI
D03	Data Science Bowl 2017	https://www.kaggle.com/c/data-science-bowl-2017/data	LDCT	-	1595	Highly imbalanced	No	DI
D04	SPIE-AAPM cases [36]	https://www.cancerimagingarchive.net/collection/spie-aapm-lung-ct-challenge/	CT	Multi	70	Nearly balanced	Yes	DI
D05	LNDb	https://zenodo.org/records/6613714	CT	Multi	294	Nearly balanced	Yes	DI
D06	DLCST	-	LDCT	Single	4104	Imbalanced	Yes	DI
D07	NLST	https://www.cancerimagingarchive.net/collection/nlst/	LDCT	Multi	26722	Highly imbalanced	No	DI
D08	KiTs19	https://kits19.grand-challenge.org/data/	CT	Multi	210	Imbalanced	No	NI
D09	TCGA-KICH	https://www.cancerimagingarchive.net/collection/tcga-kich/	CT, MR	Multi	15	Imbalanced	No	DI
D10	TCGA-KIRP	https://www.cancerimagingarchive.net/collection/tcga-kirp/	CT, MR, PT	Multi	33	Imbalanced	No	DI
D11	TCGA-KIRC	https://www.cancerimagingarchive.net/collection/tcga-kirc/	CT, MR, CR	Multi	267	Imbalanced	No	DI
D12	IQ-OTH/NCCD dataset [39]	https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset	CT	Single	110	Imbalanced	No	DI
D13	CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone [42]	https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone	CT	Single	-	Imbalanced	No	DI
D14	KAUH-kidney dataset [45]	-	CT	Single	120	Balanced	No	DI
D15	HCC-TACE-Seg	https://www.cancerimagingarchive.net/collection/hcc-tace-seg/	CT	Single	105	Imbalanced	No	DI

Table 5
Private dataset summary for CT-based lesion classification

Dataset ID	Dataset Name	Modality	Sample	Balanced/ Imbalanced	Organ	Reference
D16	AHCMU Dataset	CT	477	-	Lung	[21]
D17	Multi-center Dataset	CT/PET	647	-	Whole body	[20]
D18	NLST + LIDC-IDRI + Pluss	LDCT	1600	-	Lung	[47]
D19	Lung CT dataset	CT	1097	Imbalanced	Lung	[48]
D20	KUH dataset	CT	1177	-	Lung	[29]
D21	External in-house dataset	LDCT	47	Balanced	Lung	[9]
D22	Single institute dataset	CT	502	Imbalanced	Abdomen	[49]
D23	Fujian Cancer Hospital dataset	CT	313	Imbalanced	Lung	[23]
D24	AMDI dataset	CT	250	Imbalanced	Lung	[50]
D25	Multi-hospital dataset	CT	972	Imbalanced	Lung	[51]
D26	Multi-hospital dataset	CT	2000	Imbalanced	Liver	[52]
D27	FAHWMU dataset	CT	341	Imbalanced	Stomach	[53]
D28	Multi-center dataset	CT	864	Imbalanced	Abdomen	[54]
D29	Multi-center dataset	CT	999	-	Lung	[55]
D30	SPH dataset	CT	212	Imbalanced	Lung	[56]
D31	Multi-center dataset	LDCT	39014	-	Lung	[35]
D32	Single-center dataset	CT	1589	Imbalanced	Liver	[46]
D33	Single-center dataset	CT	502	Imbalanced	Abdomen	[57]
D34	Multi-center dataset	CT	590	Imbalanced	Lung	[58]
D35	MPCT-FLLs	CT	85	Imbalanced	Liver	[59]

3.3 Deep learning techniques in lesion classification

In response to the diverse challenges across lung and abdominal lesion classification, recent studies have explored a wide range of deep learning approaches. These approaches include custom CNNs designed for organ-specific features, pre-trained models leveraging transfer learning, hybrid networks that combine multiple learning strategies, and ensemble frameworks designed to enhance robustness and generalizability. More recently, the field has progressed toward Transformer-based models and self-supervised learning (SSL) strategies, which enhance global contextual modeling and data efficiency. Accordingly, Fig. 2 illustrates the taxonomy of DL approaches for CT lesion classification, providing a summarized view for improved clarity.

3.3.1 Custom CNN

Custom-designed CNNs have become central to lung CT lesion classification because they capture subtle morphological and textural cues that traditional methods miss [60]. Their ability to capture spatial hierarchies and handle large-scale data has led to notable success in lung nodule detection and classification [61–63]. Recent models increasingly combine lesion-focused features with broader anatomical context. The dual-pathway frameworks in [20] and [8] exemplify this shift: [20] merges detailed 3D Res-UNet features with positron emission tomography (PET) signals through 3D–2D feature fusion to distinguish lesions with diffuse boundaries or overlapping patterns. At the same time, [8] separates intra-nodule and peri-nodule cues to learn both lesion heterogeneity and its spatial relationship to vessels or pleura. Despite their differing mechanisms, both demonstrate that integrating local morphology with contextual information is more influential than simply deepening the network. However, their reliance on manual masks, high computational cost, and limited external validation continues to restrict clinical scalability.

In contrast, preprocessing-driven models such as the 15-layer CNN [48] and KNG-CNN [31] depend heavily on handcrafted filters, contrast enhancement, and GAN-generated inputs to compensate for low contrast and overlapping tissues. This creates a brittle pipeline that performs well only under controlled conditions and is highly sensitive to domain shifts. More advanced volumetric approaches, such as the 3D Faster R-CNN in [29], further highlight the value of full 3D continuity, enabling detection of faint or pleura-attached nodules. Although the model is effective as a CAD second reader, it still exhibits notable false-positive rates. Meanwhile, the multi-scale CNN in [47] shows that 3D multi-resolution modeling can improve cross-dataset generalization, though the gains are modest relative to the computational burden and annotation limitations. Recent studies have also explored lightweight custom CNN architectures with integrated explainability mechanisms. For example, a compact CNN with 182k parameters (~ 33M FLOPs) integrates SHAP (Shapley additive explanations) and LIME (local interpretable model-agnostic explanations) to highlight key regions in kidney CT lesion classification [44]. While this design demonstrates high reported accuracy and improved transparency, its limited dataset size and narrow diagnostic scope raise concerns regarding real-world generalizability. Table 6 summarizes these custom CNN approaches.

Table 6
Overview of custom CNN-based approaches for lesion classification

Model	Image Modality	Key Architectural Features	Lesion Classification Challenge Mitigated	Architectural Impact	Limitation	Reference
Dual-branch 3D CNN + 2D CNN	PET/CT	Multi-scale learning, local–systemic metabolic information integration, and 3D–2D feature fusion	overlapping patterns, diffuse boundaries	Improved lesion characterization through local–systemic metabolic information integration	Complex design complicates training	[20]
2D/3D Multi-scale CNN	LDCT	Multi-scale learning, multi-scale context representation, and multi-branch feature fusion	Varying lesion sizes	Multi-scale architecture effectively handled varying nodule sizes	High computational demand	[47]
Custom 15-layer CNN	CT	Deep feature extraction, hierarchical learning, and convolution-based representation	-	Enhanced discrimination of normal, benign, and malignant lesions	Risk of overfitting on a small dataset	[48]
Dual Pathway 3D CNN	LDCT	Target–context learning, and VAE-enhanced feature extraction	Low contrast, slight lobulation	Joint learning of intra-nodule and peri-nodule information	Limited external validation	[8]
3D Faster R-CNN	CT	3D feature extraction, lung-field extraction, region proposal network	Overlapping patterns	High sensitivity and effective detection of GGNs and small nodules.	High false positives and missed faint GGNs.	[29]

3.3.2 Pre-trained model

Pre-trained CNNs, trained initially on large-scale datasets, offer a practical solution for medical image analysis by transferring learned representations to new tasks. This approach reduces training costs and mitigates data scarcity challenges in CT imaging.

In lung CT, the MedicalNet (3D ResNet-10) [55] extends transfer learning into the volumetric domain by utilizing pretrained 3D medical image features. Its ability to learn spatially consistent volumetric representations improved discrimination between invasive adenocarcinoma (IAC) and minimally invasive adenocarcinoma (MIA), particularly in cases with overlapping imaging appearances. However, its dependence on manual ROIs, extensive preprocessing, and specialized hardware limits its practical deployment. ImageNet-pretrained architectures such as VGG, DenseNet, ResNet, and Xception [28], [40] used pretrained convolutional layers to extract hierarchical features and adapted them for CT lesion classification via partial fine-tuning. While transfer learning enabled strong performance with limited data, the observed variation across backbones suggests that classification outcomes depend not only on pretraining but also on each model's architectural capability to capture discriminative lesion features. A fully fine-tuned 50-layer ResNet [35] further improves detection of overlapping patterns and small nodules through deep residual feature learning but remains affected by incomplete clinical metadata.

Beyond lung CT, EfficientNetB5 [46] fine-tuned on real and GAN-augmented liver crops learns key enhancement cues for distinguishing visually similar lesions, though single-center data limits generalizability. VGG-16 [57], fine-tuned in its upper layers, captures liver-specific patterns while preserving generic ImageNet features. CCDA augmentation sharpens class boundaries and stabilizes low-contrast lesions. Pretrained models like Inception and ResNet have also shown promise for renal lesion classification [64–67], though performance remains sensitive to dataset quality and class imbalance. Table 7 summarizes pretrained models for lesion classification.

Table 7
Overview of pretrained models for lesion classification

Organ	Pre-trained Model	Modality	Training Strategy			Key Architectural Features	Lesion Classification Challenge Mitigation	Architectural Impact	Limitation
			Pretraining Source	Feature Extraction	Fine-tuned				
Lung	MedicalNet	CT	Large-scale 3D medical imaging datasets.	Extract volumetric spatial features from CT patches.	Partially fine-tuned.	3D MedicalNet (ResNet10) backbone and image-mask fusion	Overlapping patterns	Improved discrimination of IAC and MIA using pretrained 3D medical image features.	The model requires careful hyperparameter tuning.
Lung	VGG-16, VGG-19	CT	ImageNet	Initial feature extractor.	Partially fine-tuned.	Pretrained CNN backbone and custom classification head	-	Enhanced lesion classification through hierarchical feature representations learned by pretrained CNN backbones.	Requires pre-localized and annotated lesions, limiting end-to-end clinical applicability
	DenseNet-121			Extracted deep feature maps.					
	DenseNet-169			Extracted hierarchical features.					
	ResNet-152V2			Extracted residual feature representations.					
Lung	VGG 16	CT	ImageNet	Extracted convolutional feature maps.	-	Pretrained CNN backbones with hierarchical feature extraction	-	Xception improved classification efficiency through depth-wise separable convolutions and residual learning	VGG19 showed a higher test loss.
	VGG 19			Extracted high-dimensional deep features.					
	Xception			Extracted deep separable-convolution features.					
Lung	ResNet-50	LDCT	-	Extracted discriminative features from candidate nodule ROIs.	Fully fine-tuned	Residual connections and deep convolutional feature learning	Overlapping patterns.	Improved discrimination between nodules and non-nodules through deep residual feature learning.	Lack of a definitive gold standard and incomplete clinical metadata limit comprehensive evaluation.
Liver	EfficientNetB5	CT	ImageNet	Extracted lesion appearance features.	Yes	Balanced scaling of depth, width, and resolution for efficient feature extraction	Overlapping patterns.	Enhanced discrimination of focal liver lesion subtypes through efficient feature learning	A single-centre dataset may limit generalizability.
Liver	VGG-16	CT	ImageNet	Convolutional layers extracted texture & intensity features from 224×224 lesion patches.	Partially fine-tuned.	Deep convolutional feature learning	Overlapping patterns, diffuse boundaries.	Effective lesion representation learning with reduced overfitting risk on small datasets.	Performance relies on MixUp and AugMix augmentation.

3.3.3 Hybrid model

Hybrid CNN models integrate CNNs with complementary architectures, such as LSTMs, attention mechanisms, or radiomics, to enhance feature learning, robustness, and clinical applicability. These models aim to overcome the limitations of single-CNN pipelines by capturing richer spatial, temporal, or contextual information.

In lung CT, MSSFN [23] fuses GoogLeNet features, multi-scale convolutions, and a Bi-LSTM to distinguish PLC, CIP, and RP despite overlapping patterns and diffuse boundaries. However, it remains computationally heavy and lacks multi-center validation. PB-LNet [21] similarly pairs ResNeXt50 with a Bi-LSTM to enhance pulmonary nodule subtype classification through deep feature and sequential context learning, but its generalizability remains limited by single-center validation. CNN–LSTM hybrids such as 1D-CNN–LSTM and VGG16-1D–LSTM [50] show that temporal modeling enriches geometric cues. The LSTM–Inception model [22] strengthens this by combining detailed local features with 3D contextual relationships, improving the handling of blurred edges and subtle shapes, but increasing the risk of overfitting. Hybrid radiomics–deep frameworks [68] further enhance interpretability and robustness by integrating handcrafted texture and shape descriptors with deep contextual features, thereby improving the classification of low-contrast and overlapping lesions.

Attention-based hybrids extend this trend. CBAM-ResNet [51] refines CNN features and fuses them with clinical attributes, improving prediction for subtle nodules but remaining limited in ambiguous or low-contrast cases. In contrast, a lightweight hybrid approach such as DenseNet201–MLP [30] achieves strong classification accuracy without heavy attention or temporal modules, demonstrating that simpler hybrid architectures can still be highly effective when feature extraction and decision-making are well separated.

Hybrid frameworks have also been applied to the analysis of liver, gastric, and renal lesions. In kidney CT, segmentation–classification hybrid pipelines [38], [64] improve lesion subtype separation by integrating anatomical localization with dual-path feature extraction. However, large tumors and rare subtypes still challenge boundary accuracy and robustness. The SDC-UDA + DenseNet pipeline [53] leverages GAN-based pseudo-labeling to compensate for missing annotations but remains sensitive to pseudo-label quality. Domain-adapted and multi-scale hybrid models [19, 69, 70] consistently report high AUCs and meaningful reductions in diagnostic reading time, often outperforming junior clinicians. Recent studies have also explored privacy-preserving hybrid frameworks for distributed medical data analysis. For example, Aniso-ResCapHGBO-Net integrates ResNet-50, Capsule Networks, metaheuristic feature optimization, federated learning, and blockchain-based security to enhance spatial feature learning while preserving patient data privacy across institutions [71]. Although the model reports extremely high accuracy and strong privacy protection, the architecture is overly complex and computationally expensive. An overview of hybrid models for lesion classification is provided in Table 8.

Table 8
Overview of hybrid models for lesion classification

Organ	Hybrid Model	Architectural Role	Feature Integration Strategy	Lesion Classification Challenge Mitigation	Architectural Impact	Limitation	Reference
Lung	GoogLeNet +MSFM + Bi-LSTM	GoogLeNet: lesion feature extraction, MSFM: integrating lesion features across varying spatial dimensions, Bi-LSTM: captures slice-to-slice continuity.	Transfer learning, multi-scale spatial fusion, and sequence modelling	Low contrast, overlapping patterns, diffuse boundaries.	Improved lesion characterization by integrating local features, multi-scale context, and inter-slice dependencies.	Computationally intensive and lacks multi-center validation	[23]
Lung	ResNeXt50 + Bi-LSTM	ResNeXt50: deep lesion feature extraction and Bi-LSTM: sequential dependency learning	Transfer learning and sequential feature integration	-	Enhanced classification through deep feature and sequential context learning	Limited single-canter validation.	[21]
Lung	1D-CNN-LSTM and VGG16-1D-LSTM	1D-CNN/VGG16-1D: extracts essential feature maps, manages, LSTM: captures long-term dependencies	Regression-based handcrafted features fused with CNN/LSTM representations	-	Low-cost 1D-CNN-LSTM provided efficiency, and VGG16-LSTM achieved deeper learning with high accuracy.	Small dataset limits generalizability, while data-sharing restrictions hinder external validation.	[50]
Lung	ResNet50 + CBAM	ResNet50: deep feature extraction through residual learning; CBAM: channel and spatial attention refinement	Combines imaging with morphological and clinical features using attention	Slight lobulation	Improved lesion classification through attention-enhanced feature representation	Requires larger datasets and higher computational resources.	[51]
Stomach	SDC-UDA + DenseNet	SDC-UDA: segmentation and 3D DenseNet: volumetric feature extraction and classification	Two-stage integration of segmentation-guided ROI extraction and 3D volumetric feature learning	-	Reduced manual annotation effort, robust and repeatable results.	Limited interpretability of deep learning decision-making.	[53]
Kidney	3D Res-UNet + ResNet-50	3D Res-UNet: automated kidney and renal mass segmentation; Dual-path ResNet50: local–global feature extraction for subtype classification	Segmentation-guided ROI extraction with dual-path local–global feature fusion.	Overlapping patterns, diffuse boundaries.	Improved renal mass subtype discrimination through combined local texture and global contextual learning	Struggles with rare RCC subtypes due to class imbalance.	[38]

3.3.4 Ensembled model

Ensembling CNNs is a widely used strategy to improve robustness and reduce overfitting by combining the strengths of multiple models. ERBNet [9] employs a weighted bagging ensemble of lightweight 3D CNNs trained across various LDCT dose levels, enabling it to better handle low contrast and subtle morphological cues. However, its dependence on simulated rather than real LDCT data raises concerns about true clinical generalizability. For liver lesions, the Multi-Scale CNN Ensemble in [49] aggregates multiple ResNet-18 models trained on multi-sized LINA patches, then averages confidence scores to make a final prediction. These patches enhance faint textures and preserve boundary details, improving the discrimination of small and visually similar lesion

subtypes while reducing inter-class overlap. Yet the model's reliance on GAN-generated synthetic data introduces risks of learning artifacts that may not reflect real patient variability. A different strategy appears in the hybrid 2D + 3D CNN ensemble in [37], which integrates ResNet-50 multi-slice views with 3D Inception-v1 patches to capture both detailed volumetric cues and broader spatial context. This design improves recognition of irregular margins and performs strongly under low-dose conditions, with better external validation. Nonetheless, its reliance on single-timepoint CT scans and sample selection bias limits clinical completeness. An overview of these ensemble approaches and their trade-offs is provided in Table 9.

Table 9
Overview of ensemble models for lesion classification

Organ	Ensembled Model	Modality	Architectural Role	Feature Integration Strategy	Lesion Classification Challenge Mitigated	Architectural Impact	Limitation	Reference
Lung	ERBNet	LDCT	3D CNN (RBNet): hierarchical volumetric feature extraction and nodule classification	Performance-weighted averaging of five RBNet models trained across different dose levels (bagging-style)	Low contrast, overlapping patterns	Accurately classified lesions without pre-denoising, enhancing robustness across dose levels.	Relied on simulated LDCT data, not actual low-dose acquisitions, which may affect real-world performance.	[9]
Liver	Multi-Scale CNN Ensemble	CT	LINA patches: augment lesion texture, shape, and boundary information; ResNet-18: lesion feature extraction and classification	Bagging-style ensemble of ResNet-18 models trained on LINA patches at different scales, with class-confidence averaging	Overlapping patterns, diffuse boundaries	Enhanced discrimination of small and visually similar lesion subtypes through multi-scale ensemble feature learning	Single-institution data limits generalizability, and synthetic data may affect realism.	[49]
Lung	2D + 3D CNN	LDCT	ResNet50: multi-view 2D feature extraction; 3D Inception-v1: volumetric feature extraction for malignancy risk prediction	Hybrid ensemble fusion of multi-view 2D and volumetric 3D CNN predictions through score averaging.	-	Improved pulmonary nodule malignancy risk prediction through complementary 2D and 3D feature learning	Uses only single CT scans, ignoring longitudinal growth and selection bias from excluded cancers.	[37]

3.3.5 Transformer model

Transformer-based architectures have been introduced in CT lesion classification to overcome the locality constraint of convolutional kernels by modeling long-range spatial dependencies through self-attention.

Pure Transformer models rely primarily on patch embedding and multi-head self-attention without convolutional feature backbones. Comparative evaluations of Swin Transformer (swin-t), multi-axis vision transformer (MaxViT), and ViT-b-32 for multi-organ tumor classification showed that hierarchical swin-t achieved the most favorable accuracy–efficiency trade-off, benefiting from shifted-window attention that integrates local and global context [41]. In contrast, MaxViT achieved comparable accuracy at substantially higher computational cost, while vanilla ViT variants underperformed on limited CT datasets, reflecting weak inductive bias and strong data dependence. A Swin-based framework for pure ground-glass nodule (pGGN) subtype classification similarly demonstrated improved multi-scale contextual integration through shifted-window self-attention, enhancing discrimination of overlapping imaging patterns; however, its small retrospective cohort and lack of multi-center validation limit generalizability [58]. The multitask Swin Transformer (MTST) further extends this paradigm by incorporating task-specific branches to model radiological attributes, such as margin, lobulation, and spiculation, and outperforms CNN baselines in lung nodule classification [33]. Nevertheless, reliance on radiologist consensus labels and the absence of external validation constrain clinical reliability. Direct comparisons also reveal persistent limitations: although the data-efficient image transformer - small (DeiT-S) model benefited substantially from transfer learning, it failed to surpass a well-optimized 3D CNN baseline in lung cancer classification, underscoring optimization sensitivity and reduced robustness in small-scale settings because of the data-hungry nature of the Transformer [56].

Hybrid CNN–Transformer models combine convolutional local feature extraction with Transformer-based global context modeling to balance inductive bias and representation capacity. ViTNet integrates a ResNet50 backbone with a vision transformer (ViT) module, improving contextual understanding but showing reduced performance in low-contrast CT compared to MRI [72]. A spatio-temporal decoupling network (STD-Net) decouples spatial and temporal learning by pairing a 3D ResNet with a Transformer to model multiphasic enhancement patterns, where removal of the Transformer significantly reduced liver lesion classification performance [73]. A siamese dual-resolution transformer (SDR-Former) and multi-scale large kernel attention with mobile vision transformer (MVIT-MLKA) similarly fuse high-resolution CNN features with global Transformer attention to better handle overlapping patterns and subtle boundary variations in liver and pancreatic lesions [52], [54]. A cluster-induced mask transformer (CIMT), by combining a U-Net backbone and Transformer decoder, jointly segments and classifies gastric tumors from non-contrast CT images [74]. The model outperformed baseline models and radiologists by capturing both local texture and global contextual features, while generalizability remains limited. A broader hybrid framework combining ViT with Perceiver IO extends Transformer modeling to multimodal and multi-disease classification, achieving strong cross-domain performance but requiring substantial computational resources and lacking focused CT-specific external validation [75]. An overview of Transformer models is shown in Table 10.

Table 10
Overview of Transformer models for lesion classification

Organ	Transformer Model	Architectural Role	Feature Integration Strategy	Lesion Classification Challenge Mitigation	Architectural Impact	Limitation	Reference
Liver	SDR-Former	CNN branch: captures local/high-frequency details, Transformer branch: captures global/low-frequency context, APSM: adaptively weights and fuses multi-phase CT information	Cross-resolution fusion of CNN and Transformer features with adaptive multi-phase attention weighting	Overlapping patterns.	Improved classification performance through cross-resolution feature fusion and adaptive multi-phase attention.	Struggles with small lesions and limited-phase scenarios.	[52]
Pancreas	MVIT-MLKA	CNN: local feature extraction; Mobile ViT: global contextual modeling; MLKA: multi-scale attention refinement and local–global feature enhancement	Hybrid fusion of CNN local features and Transformer global representations with multi-scale large-kernel attention.	Low contrast, overlapping patterns, diffuse boundaries.	Outperformed standalone CNN and Transformer through complementary local–global representation learning and multi-scale attention refinement.	Required high-quality, standardized imaging data and relies on 2D slices instead of full 3D volumes.	[54]
Lung, Kidney	ViT	ViT: Global feature extraction through self-attention over image patches, swin-t: hierarchical feature extraction using localized self-attention, MaxViT: hybrid CNN–Transformer feature extraction combining local and global context	Patch embedding with global self-attention and positional encoding	Overlapping patterns	Achieved competitive classification performance with efficient feature modelling	High computational cost for the MaxViT model.	[41]
	swin-t		Shifted-window self-attention with hierarchical patch merging				
	MaxViT		MBConv convolutions with local and global attention integration				
Lung	MTST	Shared swin-t backbone: hierarchical local–global feature extraction; task-specific branches: malignancy and semantic feature prediction	Hard-parameter-sharing multitask learning with a shared swin-t encoder and task-specific prediction heads.	Low contrast, slight lobulation, overlapping patterns, diffuse boundaries.	Improved diagnostic interpretability through joint malignancy and radiological feature prediction.	Relies on radiologist-consensus labels rather than pathology-confirmed ground truth, thereby limiting biological validity.	[33]
Lung	DeiT-S	Global contextual feature extraction through self-attention	Patch embedding + positional encoding + CLS token with multi-head self-attention integration	-	Enhanced interpretability via attention maps	Failed to match the 3D CNN's performance	[56]
Lung	swin-t	Integrate multi-scale local–global features and stabilize final predictions	Hierarchical patch merging with shifted-window self-attention and cross-fold prediction integration	Slight lobulation, overlapping patterns	Efficient Swin architecture with reduced computational complexity.	High implementation and maintenance costs.	[58]

3.3.6 Self-supervised learning model

Self-supervised learning (SSL) enables models to learn meaningful representations from unlabeled medical images, reducing reliance on labor-intensive expert annotations. By capturing intrinsic structural and contextual patterns, SSL improves model initialization and downstream lesion classification performance under limited labeled data [76, 77].

Liver vision-language model (Liver-VLM) [59] performs liver lesion classification using a very small, single-center multi-phase CT dataset by adopting a phase-shuffle SSL strategy within a lightweight vision–language framework. This pretraining helps the model learn phase-aware and structural relationships before supervised fine-tuning, improving feature initialization, classification performance, and computational efficiency. However, the model remains vulnerable to small, low-contrast lesions and non-standard CT protocols, limiting its clinical robustness. Similarly, SSLSD-KTD integrates masked autoencoder (MAE)-based SSL with self-distillation (SD) to enhance kidney tumor detection (KTD) from limited label-free CT data [43]. The encoder learns local features, while the decoder captures global context and transfers this knowledge back via teacher–student distillation, improving representation quality

and reducing ViT data dependency. Although it achieves strong internal performance and outperforms MAE and ViT baselines, it still relies on supervised fine-tuning and lacks multi-center validation, limiting generalizability. A self-supervised transfer learning-based on domain adaptation (SSTL-DA) model [32] integrates SSL for robust feature extraction with adversarial domain adaptation (GRL) in a 3D framework, effectively addressing labeled data scarcity and reducing source–target discrepancies in lung nodule classification. By capturing 3D spatial heterogeneity and enhancing feature transfer, the model achieves high sensitivity and competitive performance. Nevertheless, high computational cost, background interference, and limited multi-center validation restrict clinical robustness. An overview of SSL models is presented in Table 11.

Table 11
Overview of SSL models for lesion classification

Organ	SSL Model	Architectural Role	Feature Integration Strategy	Architectural Impact	Limitation	Reference
Liver	Liver-VLM	ResNet-18: phase-aware multi-phase CT feature extraction; BERT: medical semantic encoding of lesion and phase descriptions.	Vision–language alignment through cosine-similarity matching of CT image embeddings and phase-aware medical text embeddings, supported by self-supervised phase-shuffle pretraining.	Enhanced focal liver lesion classification through phase-aware self-supervised pretraining and multimodal semantic alignment.	Struggles with low-contrast lesions, very small lesions, and non-standard scanning protocols.	[59]
Kidney	SSLSD-KTD	ViT: global feature extraction; MAE: self-supervised image reconstruction; SD: feature refinement and knowledge distillation for classification.	Self-supervised MAE pretraining combined with self-distillation (teacher–student) to enhance encoder representations before fine-tuning.	Effective on small datasets through MAE-based self-supervised learning and self-distillation.	Requires long pre-training and is partially dependent on labelled data during supervised fine-tuning.	[43]
Lung	SSTL-DA	Self-supervised 3D U-Net: transferable feature learning, domain adaptation network: domain alignment, classification network: benign–malignant prediction.	Self-supervised feature pretraining with adversarial domain adaptation and transfer learning.	Self-supervised learning reduces reliance on labelled data.	Misclassification risk from background tissue interference.	[32]

3.4 Comparing deep learning models' performance

This section provides a structured comparison of deep learning models for lesion classification in CT images. The analysis focuses on three key dimensions: model complexity, training strategies, and performance evaluation. Together, these aspects enable a comprehensive understanding of model effectiveness.

3.4.1 Architectural complexity

Architectural complexity reflects the structural depth, parameter scale, and computational requirements of DL models used for CT-based lesion classification, as detailed in Table 12. The reviewed models span a broad spectrum of architectural complexity, from shallow 2D CNNs such as AlexNet and VGG [48] to deeper pretrained networks like ResNet-50 [35], hybrid frameworks and ensemble models that combine CNNs with LSTMs, or other specialized modules [23], [21], [9], [38], [49–53], Transformer-based and SSL models [32], [33], [41], [43], [52], [54], [56], [58], [59].

Custom CNNs often achieve strong lesion-classification performance, yet many remain clinically impractical due to long training times, heavy GPU requirements, and limited multi-center validation. For example, the multi-scale CNN in [47] required ~ 13 hours to train despite having only 3.7M parameters, while dual-path and 3D architectures in [8], [20] required large GPU memory or DGX-class hardware, indicating substantial computational overhead relative to their architectural simplicity.

Pretrained CNN models, including VGG, DenseNet, ResNet-50, EfficientNetB5, and MedicalNet-based architectures [28], [35, 40, 46, 55, 57] offer a more balanced trade-off between architectural complexity and computational efficiency. Transfer learning enables these models to converge faster with fewer training iterations while maintaining strong predictive performance with standard GPUs. For example, the ResNet-50 model in LDCT screening [35] outperformed radiologists, achieving real-time inference (0.08–0.09 min per study) with 3.6×10^9 FLOPs. However, purely convolutional architectures remain limited in modeling long-range spatial relationships within volumetric CT data.

Hybrid, ensemble, and SSL models achieve excellent diagnostic accuracy but introduce substantial architectural and computational complexity. Systems such as MSSFN [23], PB-LNet [21], CBAM-ResNet [51], segmentation–classification pipelines [38], and gastric two-stage frameworks [53] rely on LSTMs, attention blocks, or multi-branch networks to capture rich 3D and multi-phase cues, often outperforming simpler CNNs and even junior clinicians. However, these gains come at the cost of high GPU requirements (RTX 3090/TITAN RTX), long training cycles, and complex inference pipelines that are difficult to run in typical clinical environments. Ensemble systems such as ERBNet [9] and multi-scale CNN ensembles [49] further increase overhead by requiring the training and maintenance of multiple models. SSL models introduce additional training complexity. SSLSD-KTD [43] uses a ViT-B/16 backbone with teacher–student self-distillation, requiring 200 pretraining and 50 fine-tuning epochs, while SSTL-DA [32] adopts a cascaded 3D U-Net architecture trained on multi-GPU systems, resulting in high computational cost despite improved representation learning.

Transformer-based models employ advanced architectures that capture global contextual relationships through self-attention mechanisms. Models such as ViT, Swin Transformer, and MaxViT utilize patch-based representations and hierarchical attention to model long-range spatial dependencies beyond the capability of conventional CNNs. However, these architectures vary in computational complexity and typically require dedicated GPU resources (e.g., RTX

2060 SUPER). For instance, ViT-b/32 contains ≈ 88.2M parameters with fast inference (~ 5.8 ms/image), whereas Swin-T offers improved efficiency with ≈ 28.3M parameters and 123 MFLOPs. In contrast, MaxViT-T requires substantially higher computation (≈ 5600 MFLOPs) due to its convolution–attention hybrid design. Similarly, hybrid transformer frameworks such as SDR-Former [52] and MVIT-MLKA combine CNN feature extraction with transformer-based contextual modeling, requiring high-end GPUs (e.g., RTX 3090 and GTX 4080) and large computational loads (e.g., 40.26 GFLOPs), reflecting their increased architectural complexity. Despite their higher computational requirements, transformer-based architectures often demonstrate superior classification performance and improved generalization across heterogeneous datasets. For instance, Swin Transformer-based models have been reported to outperform experienced radiologists [58].

In summary, architectural designs have evolved from simple CNNs to more complex hybrid, transformer, and SSL frameworks. While these architectures improve contextual feature learning and diagnostic performance, they also increase computational cost and hardware requirements, creating a trade-off between model accuracy and practical clinical feasibility.

Table 12
Overview of model complexity

Reference	Model Architecture	Hardware Platform	Complexity		
			Number of Parameters	Model Depth	FLOPs
Custom CNN					
[20]	Dual-Pathway CNN	-	-	-	-
[47]	2D/3D Multi-scale CNN	NVIDIA RTX 2070S	0.65-1.37M, 1.69-3.69M	-	1.31MFLOPs to 7.40MFLOPs
[48]	15 Layer custom CNN	-	-	15	-
[8]	Dual-Pathway 3D CNN	Nvidia DGX-2 server	-	-	-
[29]	3D Faster R-CNN	-	-	27	-
Pretrained Model					
[55]	MedicalNet	NVIDIA GeForce RTX 3090 GPU.	-	10	-
[28]	VGG-16, VGG-19, DenseNet-121, DenseNet-169, ResNet-152V2	4 GB RAM, Intel UHD Graphics 620 + NVIDIA MX150 GPU	-	16, 19, 121, 169, 152	-
[40]	VGG 16, VGG 19, Xception	Google Colab platform GPU support	-	16,19,36	-
[35]	ResNet-50	-	-	50	3.6×10^9
[46]	EfficientNet	Tesla V100 PCIe GPU with 32 GB of system RAM.	-	-	-
[57]	VGG-16	-	-	16	-
Hybrid Model					
[23]	MSSFN	NVIDIA TITAN RTX single GPU.	-	19	-
[21]	PB-LNet	-	-	4 (Bi-LSTM)	-
[50]	CNN-LSTM	-	-	18, ≈ 17+	-
[51]	ResNet50 + CBAM	-	-	50	-
[53]	Two-stage hybrid	64 GB DDR4 memory and RTX 4060 Ti graphics card.	-	121	-
[38]	3D Res-UNet + ResNet-50	NVIDIA GTX 1080ti GPU.	-	50	-
Ensemble Model					
[9]	ERBNet	-	-	-	-
[37]	2D + 3D CNN	-	-	50, 22	-
[49]	Multi-Scale CNN	-	-	-	-
Transformer Model					
[52]	SDR-Former	NVIDIA RTX 3090 GPU.	28.52M	residual blocks + attention blocks	40.26 GFLOPs
[54]	MVIT-MLKA	NVIDIA GeForce GTX 4080 super graphics card.	-	-	-
[41]	ViT, swin-t, MaxViT-t	NVIDIA GeForce RTX 2060 SUPER.	ViT: 88.19 M, Swin-t: 28.29 M, MaxViT-t: 30.92M	ViT: 12 layers	ViT: 456.96 MFLOPs, Swin-t: 123.43 MFLOPs, MaxViT-t: 5600 MFLOPs
[33]	MTST	RTX 2060 GPU	-	Swin Transformer (2–2–18–2 blocks) + classifier (4 blocks) + 8 feature branches (2 blocks each).	-

Reference	Model Architecture	Hardware Platform	Complexity		
			Number of Parameters	Model Depth	FLOPs
[56]	DeiT-S	TITAN RTX GPUs	-	-	-
[58]	swin-t	NVIDIA RTX TITAN	-	-	-
SSL Model					
[59]	Liver-VLM	NVIDIA RTX A6000	-	Dual-encoder + FC projection layers	-
[43]	SSLSD-KTD	-	-	ViT-B/16 backbone + student–teacher self-distillation with MLP heads	-
[32]	SSTL-DA	6 NVIDIA RTX TITAN	-	3D U-Net encoder–decoder + GAP + 3 FC layers	-

3.4.2 Model training strategies and optimization configurations

Model training plays a crucial role in enhancing the performance and generalizability of deep learning models for lesion classification from CT images, as detailed in Table 13.

Augmentation

Data scarcity, imbalance, and overfitting, a recurring challenge in CT imaging, make augmentation and synthetic data generation essential. The reviewed studies employed three main categories of augmentation techniques: transformation-based, generative, and distortion-based augmentation.

Transformation-based augmentation was the most commonly used approach and included rotation, flipping, scaling, cropping, affine transformation, brightness adjustment, contrast adjustment, mirroring, and phase-shuffle permutation [8], [29], [32], [33], [38, 55], [43], [47], [52], [54], [56], [58], [59]. These techniques increase training data diversity while preserving lesion characteristics, thereby improving robustness and reducing overfitting. In addition, class-aware augmentation strategies such as CCDA [57] combined MixUp and AugMix according to class distribution, demonstrating improved minority-class representation and sensitivity in highly imbalanced datasets.

Generative augmentation was primarily employed to address limited data availability and class imbalance through synthetic sample generation using GAN-based approaches. Study [48] utilized GANs to generate synthetic benign, malignant, and normal CT images, improving dataset balance and model generalizability. Similarly, [46] employed class-specific DCGAN models to generate realistic focal liver lesion images, enhancing data diversity while preserving lesion-specific characteristics. In [33], U-Net GAN was used to synthesize pulmonary nodule images and balance malignancy classes, resulting in improved classification performance compared with conventional augmentation alone. Unlike image synthesis for class balancing, [53] incorporated CycleGAN within an unsupervised domain adaptation framework to generate pseudo-labeled portal-phase images, addressing the scarcity of annotated data. Furthermore, [49] proposed a two-stage DCGAN–pix2pix framework that generated both lesion masks and corresponding CT images, reducing feature overlap between lesion categories and improving class discrimination.

Distortion-based augmentation introduces controlled perturbations to improve robustness against image noise, acquisition variability, and anatomical heterogeneity. Gaussian noise was employed in [29] to simulate low-dose CT conditions, enabling stable detection performance across different radiation dose levels, while [47] incorporated Gaussian additive noise during online augmentation to reduce overfitting and improve model reliability under varying imaging conditions. To address anatomical variability, [21] applied elastic distortion to expand limited and imbalanced datasets while preserving lesion characteristics, whereas [38] utilized elastic deformation to improve generalization to diverse tumor and kidney morphologies. Additionally, [9] introduced Poisson noise to simulate realistic low-dose CT images, exposing the model to varying image quality and noise levels encountered in clinical screening.

Furthermore, to mitigate severe class imbalance, the synthetic minority oversampling technique (SMOTE) has been applied to oversample the underrepresented category, ensuring more balanced training data [8].

Transfer learning

Transfer learning (TL) was one of the most widely adopted training strategies for improving lesion classification performance [21], [23], [28], [32], [33], [40], [41], [43], [48], [49], [51], [52], [55], [56], [35], [46], [57], [58], [59], particularly when annotated medical datasets were limited. By leveraging pretrained representations from large-scale datasets such as ImageNet or medical-domain pretraining frameworks, TL enabled models to learn more discriminative features and adapt effectively to CT-based lesion classification tasks. Several studies reported improvements in feature extraction and classification performance through TL. For example, [23] demonstrated that TL enhanced preliminary feature extraction while suppressing background noise, whereas [55] showed that MedicalNet-based pretraining provided strong 3D feature representations for pulmonary nodule analysis. Similarly, [32] reported improved feature separability between benign and malignant nodules, while [21] utilized pretrained ResNeXt50 features to improve classification accuracy. Where direct comparisons were performed, TL consistently outperformed random initialization or training from scratch. For instance, [56] reported that TL increased the DeiT-S AUC from 89.2% to 95.3%, while [43] demonstrated substantial performance gains on both kidney tumor datasets after incorporating pretrained knowledge. Likewise, [32] achieved improvements in accuracy, sensitivity, specificity, and AUC compared with a model trained from scratch, and [52] showed

that cross-modality TL improved classification performance across CT and MR datasets. Furthermore, several studies adopted TL to mitigate overfitting and improve model adaptation when training data were limited [41], [55], [57]. Collectively, these findings indicate that TL improves feature representation learning and data efficiency, making it one of the most effective strategies for enhancing lesion classification performance under limited-data conditions.

Optimization configurations

Optimization configurations determine how model parameters are updated during training and therefore influence convergence stability, training efficiency, and final classification performance.

Loss function

A core component of the training process is selecting an appropriate loss function, which directly influences how well the model learns to minimize error. As shown in Table XIII, cross-entropy-based losses dominate lesion classification tasks due to their effectiveness in learning probabilistic predictions. For multi-class classification, categorical cross-entropy (CCE) was the most frequently adopted loss [20], [21], [23], [38], [41], [48], as it effectively guides learning by penalizing low probabilities assigned to the correct class. In contrast, for binary classification tasks, binary cross-entropy (BCE) is preferred, as it optimally handles two-class scenarios by encouraging the model to assign high confidence to the correct category [28], [9], [8], [50]. Additionally, several studies further incorporated specialized loss functions to address specific challenges. For example, focal loss [55] was used to mitigate class imbalance by emphasizing hard-to-classify samples, while the combination of cross-entropy and Dice loss [38] improved both lesion localization and classification performance. Similarly, self-supervised frameworks employed reconstruction-based objectives, such as mean squared error (MSE) and self-distillation losses [43] to learn robust feature representations from unlabeled data. The choice of loss function can substantially influence model convergence, classification performance, and robustness.

Optimizer

Another key component of model training is the optimizer, as it determines how network parameters are updated to minimize the loss function. The Adam optimizer remains a popular default due to its adaptive learning rate capabilities and stable convergence behavior in medical image classification tasks [8, 9, 23, 32, 33, 38, 40, 41, 47, 49–51, 55, 56, 59]. The stochastic gradient descent (SGD) is frequently employed, particularly in CNN-based architectures, owing to its simplicity and effectiveness in large-scale training [20], [48], [50], [55]. In addition, the RMSprop optimizer [28], [50] is often employed for handling non-stationary or noisy gradients, as it adaptively adjusts learning rates using a moving average of squared gradients, thereby improving training stability. More recently, the AdamW optimizer [52], [54], [56], [58] has gained attention and popularity in Transformer-based models because its decoupled weight decay mechanism provides better regularization and reduces overfitting.

Learning rate

The learning rate (LR) is a critical hyperparameter that influences optimization efficiency and model convergence. Three main strategies were identified across the reviewed studies: fixed learning rates, adaptive optimizer-based learning rates, and dynamic learning-rate scheduling. Fixed learning rates were employed in several studies [21], [48], [49], [46], [57], where a constant value was maintained throughout training. In contrast, many studies adopted adaptive optimizers such as Adam, RMSprop, and AdamW [8], [28], [33], [9], [50], [51], [54], [59], which automatically adjust parameter updates during optimization. More advanced approaches incorporated dynamic scheduling strategies, including one-cycle scheduling [20], cosine annealing [23], [52], [56], warm-up scheduling [52], [56], warm restarts [55], step decay [32], [38], [53], exponential decay [58], and plateau-based reduction [47]. These strategies progressively modify the LR during training to balance rapid initial learning with finer optimization in later stages. Overall, the growing adoption of dynamic learning-rate scheduling in transformer-, hybrid-, and self-supervised learning-based models suggests its importance for achieving efficient optimization and stable convergence in CT lesion classification tasks.

Other optimization strategies

Additional optimization strategies include batch size tuning, early stopping, and regularization techniques such as dropout, weight decay, and label smoothing [8, 20, 21, 23, 28, 29, 40, 47, 48]. These strategies were primarily employed to reduce overfitting, improve training stability, and enhance model generalization, particularly when training on limited or heterogeneous CT datasets.

Table 13
Overview of model training strategy and optimization configurations

Reference	Model Architecture	Optimizer	Loss Function	Batch Size	Learning Rate	Epoch	Transfer Learning	Augmentation		
								Transformation	Generative	Distortion
Custom CNN										
[20]	Dual-Pathway CNN	SGD	CE Loss	128	Dynamic scheduling	1500	×	-	-	-
[47]	2D/3D Multi-scale CNN	Adam	-	-	Dynamic scheduling	-	×	Rotations, affine transforms, intensity shift	-	Gaussian noise
[48]	15 Layer custom CNN	SGD	CE Loss	32	Fixed: 0.001	10	√	-	GANs	-
[8]	Dual-Pathway 3D CNN	Adam	BCE loss,	-	Adaptive optimizer	150	×	3D flips, affine transforms	SMOTE	-
[29]	3D Faster R-CNN	-	-	-	-	-	×	Rotation, scaling	-	Gaussian noise
Pretrained Model										
[55]	MedicalNet	Adam, SGD	Focal loss	16	Dynamic scheduling	-	√	Flipping	-	-
[28]	VGG-16, VGG-19, DenseNet-121, DenseNet-169, ResNet-152V2	-	BCE loss	32	Adaptive optimizer	50	√	Keras Image Data Generator	-	-
[40]	VGG 16, VGG 19, Xception	Adam	CCE loss.	32	-	20	√	Keras Image Data Generator	-	-
[35]	ResNet-50	-	-	-	-	305	√	-	-	-
[46]	EfficientNet	Adam	BCE loss	128	Fixed: 0.0008	30	√	Conventional	DCGAN	-
[57]	VGG-16	-	-	8	Fixed: 0.0002	100	√	CCDA	-	-
Hybrid Model										
[23]	MSSFN	Adam	CE loss	16	Dynamic scheduling	50	√	-	-	-
[21]	PB-LNet	-	CE loss	8	Fixed: 0.01	-	√	-	-	Elastic distortion
[50]	CNN-LSTM	Adam, SGD, RMSprop	BCE loss	128	Adaptive optimizer	400	×	-	-	-
[51]	ResNet50 + CBAM	Adam	CE loss	128	Adaptive optimizer	-	√	-	-	-
[53]	Two-stage hybrid	-	-	2, 20	Dynamic scheduling	100, 200, 1000	×	-	CycleGAN	-
[38]	3D Res-UNet + ResNet-50	Adam	CE+Dice loss, weighted CE loss	2, 32	Dynamic scheduling	300, 100	×	Random rotations, scaling, gamma correction, and mirroring	-	Elastic deformatio
Ensemble Model										
[9]	ERBNet	Adam	BCE loss	-	Adaptive optimizer	40	×	-	-	Poisson noise
[37]	2D + 3D CNN	-	-	-	-	-	×	-	-	-

Reference	Model Architecture	Optimizer	Loss Function	Batch Size	Learning Rate	Epoch	Transfer Learning	Augmentation		
								Transformation	Generative	Distortion
[49]	Multi-Scale CNN	Adam	-	20	Fixed: 0.0001	10	√	Rotation/scaling	Two-stage GAN	-
Transformer Model										
[52]	SDR-Former	AdamW	CE loss	8	Dynamic scheduling	200	√	rotations, erasing, and flips	-	-
[54]	MVIT-MLKA	AdamW	-	16	Adaptive optimizer	200	×	Random crop, horizontal flip	-	-
[41]	ViT, swin-t, MaxViT	Adam	CCE loss	4–32	-	5-200	√	-	-	-
[33]	MTST	Adam	CE loss	8	Adaptive optimizer	50	√	Flipping, rotation, brightness/contrast adjustment, and scaling	U-Net GAN	-
[56]	DeiT-S	Adam, AdamW	-	8, 16	Dynamic scheduling	300, 500	√	Random anisotropy, vertical and horizontal flip, blur, and affine	-	-
[58]	swin-t	AdamW	Weighted BCE loss	4	Dynamic scheduling	100	√	Random flipping and rotation	-	-
SSL Model										
[59]	Liver-VLM	Adam	CE loss	32	Adaptive optimizer	200	√	Phase-shuffle (permutation-based)	-	-
[43]	SSLSD-KTD	-	MSE reconstruction loss + CE self-distillation loss	-	-	50, 200	√	Random resized cropping and patch masking	-	-
[32]	SSTL-DA	Adam	CE loss	32	Dynamic scheduling	-	√	Rotation, horizontal flip	-	-

3.4.3 Performance and objective evaluation

Performance across the reviewed CT-based lesion classification models varied widely depending on model design, dataset characteristics, and evaluation setting. Several evaluation metrics are reported across the studies, but how these metrics translated to real-world reliability differed substantially. As summarized in Table 14, the reviewed studies reported a range of outcomes.

Performance and generalizability analysis

Performance reflects a model's ability to accurately classify lesions, whereas generalizability refers to its ability to maintain predictive performance across independent datasets, institutions, imaging protocols, and clinical populations. In this review, generalizability was assessed using multiple indicators, including external validation, dataset diversity, cross-validation, and clinically relevant reader studies.

Internal performance was often high, especially for custom CNNs and hybrid models. For example, the 15-layer CNN in [48] achieved a near-perfect accuracy of 0.994 and an F1 of 0.999 with GAN augmentation. Similarly, CNN-LSTM hybrids in [50] achieved 100% sensitivity, indicating a strong ability to detect malignant lesions without false negatives. Hybrid models such as MSSFN [23], PB-LNet [21], and SDR-Former [52] consistently delivered high AUCs (0.90–0.96), showing that integrating temporal, multi-scale, or transformer features improves detection of subtle or overlapping patterns. Ensemble systems like ERBNet [9] reached one of the highest reported AUCs (0.981), demonstrating the power of multi-learner fusion. SSL models such as SSLSD-KTD [43] and SSTL-DA [32] provide strong performance on small unlabeled datasets. However, strong internal datasets performance rarely translates into external datasets. For instance, the dual-pathway CNN [20] fell from 0.884 AUC internal testing to 0.807 and 0.666 on two external datasets, with sensitivity declining to 0.383, revealing severe overfitting. Even pretrained models showed similar patterns: MedicalNet [55] decreased from 0.92 to 0.82 AUC on external validation, while ERBNet [9] dropped ~ 9–10% across hospital sites. These declines indicate that many CNN-based models learn dataset-specific patterns rather than generalizable lesion characteristics. On the contrary, Transformer-based models demonstrate improved external performance. For example, Swin Transformer [58] achieved approximately 7% higher performance across evaluation metrics on independent multi-institutional cohorts' external data, while MVIT-MLKA [54] showed about a 3% improvement in sensitivity in an independent external centre dataset, suggesting improved generalization compared with many CNN-based models.

Dataset diversity also varied considerably across studies. Several studies utilized multi-center datasets [20], [51], [52], [54], [55], [35], [58], which expose models to greater variations in patient populations, imaging protocols, and scanner characteristics. In contrast, many studies relied on single-center cohorts [21], [23], [48], [49], [50], [56], [46], [57], [59], increasing the risk of selection bias and learning site-specific imaging characteristics. For example, PB-LNet [21] was developed using surgically confirmed cases from a single institution, while the two-stage hybrid model [53] and VGG-16 [57] were trained under

institution-specific imaging protocols. Such settings may limit model applicability across different patient populations, scanners, and clinical environments. In addition, when some studies employed cross-validation [8], [28], [32], [9], [38], [54], [58], most relied on a single train-test split. The limited use of cross-validation may reduce confidence in reported performance, as models are evaluated on fewer data partitions and may be more susceptible to dataset-specific bias. Furthermore, only a limited number of studies evaluated model performance in clinically relevant settings through reader studies or comparisons with radiologists [29], [35], [37], [46] [54], [58]. These findings suggest that, despite strong internal performance, the generalizability of many CT-based lesion classification models remains insufficiently validated.

Figure 3 presents a vertical bar chart comparing the AUC values of different DL models, including custom CNN, pretrained CNN, hybrid, ensemble, Transformer-based, and SSL models applied to CT-based lesion classification. Among the included studies, Transformer-based models, particularly Swin-T, MaxViT, and ViT, were reported with some of the highest AUC values. SSL, hybrid, and ensemble models also demonstrate competitive performance, while pretrained CNNs show relatively consistent results. In contrast, custom CNN architectures, such as the dual-pathway CNN, exhibit comparatively lower AUC values. These comparisons provide a descriptive overview of the reported performance and should be interpreted cautiously because the included studies differed in datasets, lesion types, classification tasks, sample sizes, and evaluation protocols.

Figure 4. Radar chart comparing the average performance of six different DL model categories across six evaluation metrics (AUC, accuracy, sensitivity, specificity, precision, and F1-score). The comparison shows that transformer-based models consistently achieve the highest values across most metrics, indicating strong and balanced predictive performance relative to other DL architectures. These average values were calculated from the reported results of the included studies and are intended to provide a descriptive overview of performance trends rather than pooled estimates from a meta-analysis.

Evaluation metric analysis

The reviewed studies primarily evaluated performance using AUC, accuracy, sensitivity, specificity, precision, and F1-score. However, these metrics often present different perspectives on model effectiveness. Accuracy values varied considerably across the reviewed studies, ranging from 0.415 on an external cohort [20] to 0.993 for a custom CNN [48]. However, high overall accuracy did not always correspond to balanced lesion classification. In the imbalanced dataset ensemble model [49], achieved an overall accuracy of 0.873, yet sensitivity remained substantially lower for hemangiomas (0.667) than for cysts (0.867) and metastases (0.920). Similarly, VGG16 [57] showed an overall accuracy of 0.700, accompanied by lower sensitivity (0.644) and a minority-class F1-score of 0.397, demonstrating inconsistencies between overall accuracy and class-wise performance. AUCs also ranged widely, 0.981 in [9] vs. 0.666 in [20], indicating dependence on dataset characteristics. Despite achieving high AUC scores (0.884, 0.920, 0.86, and 0.936) in studies [20], [47], [49], [54], they reported relatively low sensitivity of 0.566 [20], 0.774 [47], 0.667 [49], and low specificity of 0.781 [54], suggesting that a high AUC alone does not guarantee balanced diagnostic performance. Sensitivity–specificity trade-offs were frequently observed across the reviewed studies. For example, MVIT-MLKA [54] achieved high sensitivity (0.954) but comparatively lower specificity (0.726), whereas VGG-16 [57] reported higher specificity (0.841) at the expense of lower sensitivity (0.644). This illustrates the difficulty of simultaneously optimizing both metrics. This illustrates the difficulty of simultaneously optimizing both metrics. Precision and F1 varied sharply from 0.999 with GAN augmentation [48] to 0.36–0.44 without it [20], highlighting instability across conditions.

Model performance significance was evaluated in several studies using statistical analyses such as DeLong tests [8], [47], [54], [59], McNemar tests [32], [54], permutation tests [8], [38], t-tests [29], [9], [37], [52], [57], [54], [59], and JAFROC analysis [29]. However, many studies reported performance improvements without formal significance testing.

Beyond quantitative metrics, several studies incorporated interpretability techniques such as gradient-weighted class activation mapping (Grad-CAM) and class activation mapping (CAM) to visualize image regions contributing to model predictions [21, 43, 46, 51–54, 58, 59]. However, interpretability remained limited. For example, [53] reported that Grad-CAM identified where the model focused but did not reveal the specific morphological features underlying its decisions, while [54] and [46] noted that further improvements in explainability are required to enhance clinician trust and address the black-box nature of deep learning models.

Overall, these findings show that isolated metrics can be misleading, and true clinical reliability requires consistent validation across diverse datasets, institutions, and evaluation settings.

Table 14
Performance comparison of deep learning models for CT-based lesion classification

Reference	Model	Dataset Reference	Split	AUC	Acc	Sen/ Recall	Spec	Precision	F1-Score
Custom CNN									
[20]	Dual-Pathway CNN	D17 Internal	Train/Val/Test: 4:1:1	0.884	0.677	0.566	-	0.577	0.558
		external I	Test	0.807	0.569	0.427		0.440	0.405
		external II		0.666	0.415	0.383		0.441	0.365
[47]	2D/3D Multi-scale CNN	D18	Train/Val/Test: 8:1:1	3D CNN: 0.954, 2D CNN: 0.920	3D CNN: 0.886, 2D CNN: 0.827	3D CNN: 0.899, 2D CNN: 0.774	3D CNN: 0.873, 2D CNN: 0.882	-	-
[48]	15 Layers Custom CNN	D19	Train/Test: 7:3	-	0.993	0.999	-	0.999	0.999
[8]	Dual-Pathway 3D CNN	D03	Train/Val/Test: 7:1:2	0.936	-	-	-	-	-
[29]	3D Faster R-CNN	D01, D04, D05, D20	Train/Val: D01 + D20; Test: D04, D05, D20.	-	-	Sensitivity held steady at 0.960 across all SD levels (10–25).	-	-	-
Pretrained Model									
[55]	MedicalNet	D29 Internal	train/internal Val 8:2;	0.920	0.880	0.840	0.930	-	0.880
		External I	external Val.	0.890	0.830	0.830	0.830		0.830
		External II		0.820	0.780	0.800	0.770		0.780
[28]	VGG-16	D01	Train/Val (882): 8:2, Test: 125	0.954	0.871	0.821	0.921	-	-
	VGG-19			0.957	0.871	0.855	0.901		
	DenseNet-121			0.945	0.904	0.871	0.936		
	DenseNet-169			0.940	0.890	0.903	0.870		
	ResNet-152V2			0.946	0.880	0.839	0.936		
[40]	VGG 16	D30	Train/test: 7.5: 2.5	0.939	0.834	0.832	-	0.836	0.834
	VGG 19			0.926	0.809	0.809		0.811	0.810
	Xception			0.934	0.897	0.897		0.897	0.897
[35]	ResNet-50	D31	Train: 38,414; Val: 600	0.860	-	0.730	0.850	-	-
		D02	External test: 888	0.900		0.820	0.820		
[46]	EfficientNetB5	D32	Train/test: 8:2	With GAN: 0.980; without GAN: 0.950	With GAN: 0.970;	With GAN: 0.960; without GAN: 0.920	With GAN: 0.980, without GAN: 0.960	-	With GAN: 0.960;

Note: Val= validation, Acc= accuracy, AUC= area under the curve, Sen= sensitivity, Spec= specificity

Reference	Model	Dataset Reference	Split	AUC	Acc	Sen/ Recall	Spec	Precision	F1-Score
					without GAN: 0.920				without GAN: 0.920
		D32		With GAN: 0.980; without GAN: 0.970	With GAN: 0.980; without GAN: 0.960	With GAN: 0.980; without GAN: 0.960	With GAN: 0.980; without GAN: 0.960		With GAN: 0.980; without GAN: 0.960
		D15 External		-	0.982	0.922	0.98		0.875
[57]	VGG-16	D33	Train/Val/test: 5.3:2.3:2.4	-	0.700	0.644	0.841	-	0.622 Hemangioma 0.397
Hybrid Model									
[23]	MSSFN	D23	Train/Val: 8:2	0.901	0.825	0.920	-	0.842	0.92
[21]	PB-Lnet	D16	Train/Val/test: 6:2:2	-	0.84, On real data: 89%	-	-	-	-
[50]	1D-CNN-LSTM	D24	Train/Val: 7:3	0.910	0.960	0.947	-	0.900	0.923
	VGG16-1D-LSTM			0.820	0.960	1.000		0.850	0.918
[51]	ResNet + CBAM	D25	Train/test: 8:2	0.957	0.898	0.971	0.864	-	-
[53]	Two-stage hybrid	D27	Train/Val: 8:2	0.96	0.92	0.90	0.89	-	0.93
[38]	3D Res-Unet + ResNet-50	D08	Train/Val	Clear cell: 0.871, papillary: 0.862, chromophobe: 0.859, oncocytoma: -0.799	-	Clear cell: 0.841, papillary: 0.824, chromophobe: 0.837, oncocytoma: 0.769	Clear cell: 0.843, papillary: 0.837, chromophobe: 0.821, oncocytoma: 0.777	-	-
		D09, D10, D11	External Val						
Ensemble Model									
[9]	ERBNet	D02	Train: 1200, Internal test: 800	0.981	0.950	0.955	0.86	0.945	0.949
		D21	External test: 120	0.854	0.855	0.900	0.816	0.83	0.864
[37]	2D + 3D CNN	D07	Train/Val	0.91	-	-	-	-	-
		D06, Subset A,	External Val	D06: 0.93, Subset A:					

Note: Val= validation, Acc= accuracy, AUC= area under the curve, Sen= sensitivity, Spec= specificity

Reference	Model	Dataset Reference	Split	AUC	Acc	Sen/ Recall	Spec	Precision	F1-Score
		Subset B		0.96, Subset B: 0.86					
[49]	Multi-Scale CNN	D22	Train: 681, Val: 302, Test: 307	Cyst: 0.956, Hemangioma: 0.861, Metastasis: 0.937	0.873	Cyst: 0.867, Hemangioma: 0.667, Metastasis: 0.919	Cyst: 0.877, Hemangioma: 0.895, Metastasis: 0.829	-	-
Transformer Model									
[52]	SDR-Former	D26	Train/Val/test: 7.5:1:1.5	0.916	0.853	-	-	-	0.852
[54]	MVIT-MLKA	D28 Internal dataset	Train/internal test: 8:2	0.935	0.866	0.928	0.781	-	-
		External dataset	External Val	0.924	0.874	0.954	0.726		
[41]	VisT, swin-t, MaxViT	D12	Train/Val/Test: 8:1:1	ViT: 0.982 swin-t: 0.993 MaxViT-t: 0.979	ViT: 0.952 win-t: 0.971 MaxViT-t: 0.942	ViT: 0.879 swin-t: 0.917 MaxViT-t: 0.874	-	ViT: 0.965 swin-t: 0.974 MaxViT-t: 0.898	ViT: 0.911 swin-t: 0.939 MaxViT-t: 0.883
		D13		ViT: 0.997 swin-t: 1.000 MaxViT-t: 1.000	ViT: 0.977 swin-t: 0.997 MaxViT-t: 0.997	ViT: 0.964 swin-t: 0.996 MaxViT-t: 0.995		ViT: 0.983 swin-t: 0.997 MaxViT-t: 0.998	ViT: 0.972 swin-t: 0.996 MaxViT-t: 0.996
[33]	MTST	D01	Train/Val/Test: 8:1.5: 0.5	0.985	0.937	0.915	0.961	-	
[56]	DeiT-S	D30	Train/Val: 7:3	Without TL: 0.892	Without TL: 0.812	Without TL: 0.817	-	Without TL: 0.805	Without TL: 0.808
				With TL: 0.953	With TL: 0.875	With TL: 0.862		With TL: 0.874	With TL: 0.867
[58]	swin-t	D34 Internal dataset	Train/Test: 7:3	-	0.848	0.848	-	0.851	0.848
		External dataset	External Val		0.914	0.914		0.919	0.914
SSL Model									
[59]	Liver-VLM	D32	-	0.940	0.856	-	-	-	-
[43]	SSLSD-KTD	D14	Train/Test: 8:2	-	0.980	0.980	0.981	0.981	0.980
		D13			0.821	0.823	0.819	0.823	0.823

Note: Val= validation, Acc= accuracy, AUC= area under the curve, Sen= sensitivity, Spec= specificity

Reference	Model	Dataset Reference	Split	AUC	Acc	Sen/ Recall	Spec	Precision	F1-Score
[32]	SSTL-DA	D01	-	Without TL + DA: 0.944	Without TL + DA: 0.864	Without TL + DA: 0.844	Without TL + DA: 0.884	With TL + DA: 0.911	With TL + DA: 0.910
				With TL: 0.956	With TL: 0.903	With TL: 0.899	With TL: 0.906		
				With TL + DA: 0.958	With TL + DA: 0.910	With TL + DA: 0.909	With TL + DA: 0.912		

Note: Val= validation, Acc= accuracy, AUC= area under the curve, Sen= sensitivity, Spec= specificity

4 Discussion

This review systematically examined DL approaches for lesion classification from CT images, focusing on model design, training strategies, and performance evaluation. The findings provide several insights into both the progress achieved and the challenges that remain in this field.

4.1 Deep learning models and organ-specific trends (RQ1)

The reviewed studies demonstrate a clear evolution in the adoption of deep learning architectures for CT-based lesion classification. While CNN-based approaches remain the foundation of most classification systems, recent studies have increasingly shifted toward hybrid and Transformer-based architectures. These models were applied across multiple organs, including the lung, liver, stomach, kidney, and pancreas, reflecting growing interest in combining local feature extraction with global contextual learning for a more robust representation of heterogeneous lesion patterns.

When broken down by organ, distinct preferences emerge. Custom-designed CNNs are primarily applied to lung lesions, supported by the availability of public benchmark datasets, such as LIDC-IDRI, Data Science Bowl 2017, SPIE-AAPM, and LNDb, which have enabled the development and benchmarking of task-specific architectures for pulmonary nodules. Pretrained CNNs (e.g., VGG, ResNet, EfficientNet) and SSL models are widely adopted in lung, liver, and kidney lesion classification. Transfer learning and SSL strategies help to mitigate data scarcity by leveraging prior knowledge or unlabeled data to improve feature representation. Ensemble models are particularly common in lung and liver studies, where combining multiple classifiers helps balance sensitivity and specificity in challenging diagnostic settings.

A general pattern is that organs with larger datasets and strong structural features (e.g., lungs) favor custom or ensemble designs. In contrast, organs with smaller, weakly annotated, and less balanced datasets (e.g., liver, kidney, pancreas) often rely on pretrained and SSL models. Meanwhile, hybrid and Transformer-based models dominate across all organs, highlighting a broader research trend toward architectures that integrate multiple feature-extraction strategies with long-range contextual dependencies to improve robustness and generalizability. Notably, ResNet appears as a backbone across all categories of architectures (custom CNNs, pretrained CNNs, hybrids, ensembles, Transformer-based hybrid, and SSL), making it one of the most consistently applied models in CT-based lesion classification. This shows that ResNet's flexible depth, residual connections, and general robustness make it a go-to-choice for researchers, whether used directly as a pretrained model, modified into custom architectures, or integrated into hybrid/ensemble/SSL frameworks

4.2 Influence of architectural choices and training strategies (RQ2 & RQ3)

Architectural design strongly shaped model performance and generalizability. Earlier CNN-based architectures demonstrated strong performance by effectively extracting local morphological and textural characteristics. These local representations are particularly suitable for lesion-level classification, where subtle structural variations often distinguish benign from malignant lesions. However, conventional CNNs possess inherently limited receptive fields, restricting their ability to capture long-range spatial dependencies and broader anatomical context. Consequently, although several CNN-based models achieved excellent performance on internal datasets, their performance frequently declined during external validation, suggesting that they learned dataset-specific imaging characteristics rather than disease-related features. This limitation was evident in the Dual-Pathway CNN [20] and MedicalNet [55], both of which demonstrated notable performance degradation during external validation despite strong internal results, which limit their generalizability.

In contrast, Transformer-based architectures extended contextual learning through self-attention mechanisms, allowing direct modeling of long-range spatial relationships that are difficult for conventional CNNs to capture efficiently. This capability enables the model to simultaneously consider lesion characteristics and their surrounding anatomical environment, contributing to more balanced performance across multiple evaluation metrics. Accordingly, Swin Transformer [58] and Vision Transformer-based models consistently achieved among the highest AUC values reported in the reviewed studies while demonstrating relatively stable performance on independent datasets. Nevertheless, these advantages were not universal. Transformer architectures generally required substantially larger training datasets because they lack the strong local inductive bias inherent to convolutional networks. Consequently, DeiT-S [56] failed to outperform a well-optimized 3D CNN under limited-data conditions, indicating that architectural sophistication alone cannot compensate for insufficient training data.

Across the reviewed studies, the strongest and most consistent performance was achieved not by replacing CNNs with Transformers but by integrating complementary architectural components. Recent studies increasingly combined CNNs with complementary modules rather than relying on either architecture alone. Hybrid architectures integrated attention mechanisms, LSTMs, segmentation networks, or Transformer modules to model local lesion

characteristics and global contextual information jointly. Consistent performance improvements reported by MSSFN [23], PB-LNet [21], SDR-Former [52], MVIT-MLKA [54], and ResNet-CBAM [51] suggest that feature fusion, contextual modeling, and attention mechanisms improve discrimination of low-contrast lesions, overlapping boundaries, and heterogeneous lesion morphology effectively by enabling the model to emphasize diagnostically relevant regions while preserving contextual information. These findings suggest that recent performance gains are primarily attributable to architectural complementarity, in which different model components learn complementary image features that improve classification performance.

SSL further enhanced architectural robustness by learning transferable feature representations from large volumes of unlabeled CT images before supervised fine-tuning. This strategy improved performance and generalizability in data-limited settings [32], [43], reducing dependence on extensive manual annotation while complementing existing CNN and Transformer architectures. Pretrained CNNs also remained highly competitive. Models such as ResNet-50 [35] and EfficientNet [46] combined strong classification performance with efficient inference and stable external validation, indicating that carefully optimized pretrained architectures remain practical alternatives to more computationally intensive hybrid and Transformer models.

Training strategies played an equally important role in improving classification performance. Among the reviewed approaches, data augmentation and transfer learning emerged as the two most consistently adopted strategies, reflecting the widespread challenges of limited annotated data, class imbalance, and overfitting in CT-based lesion classification. Rather than simply increasing the amount of training data, advanced augmentation techniques, including GAN-based image synthesis, class-aware augmentation, elastic deformation, and noise simulation, improved dataset diversity while preserving clinically relevant lesion characteristics. These approaches were particularly effective for enhancing minority-class representation and reducing model bias toward majority classes, resulting in more robust classification across heterogeneous lesion distributions. Transfer learning similarly proved to be an effective strategy for improving feature representation under limited-data conditions. By leveraging pretrained knowledge from large-scale natural-image or medical-image datasets, transfer learning consistently improved feature discrimination, accelerated model convergence, and reduced overfitting compared with training from scratch [32].

Overall, the evidence suggests that no single architecture is universally optimal for CT-based lesion classification. Rather, model performance depends on the interaction between architectural design, dataset characteristics, and training strategy. CNNs remain effective for learning local lesion morphology, Transformers provide superior contextual modeling when sufficient training data are available, and hybrid architectures offer a balanced integration of both capabilities. Future research should therefore prioritize architectures that are not only accurate on internal benchmarks but also robust across diverse clinical datasets through external validation, standardized evaluation protocols, and clinically representative patient populations.

4.3 Evaluation metrics challenges and clinical relevance (RQ4)

CT-based lesion classification models are commonly evaluated using quantitative metrics such as accuracy, AUC, sensitivity, specificity, precision, and F1-score. Despite extensive use of quantitative metrics, their ability to fully capture clinical utility remains limited. High performance on paper often fails to translate into reliable decision-making, particularly in imbalanced datasets where rare but critical lesions may be underrepresented. Accuracy was the most frequently reported metric; however, its reliability is questionable in the predominantly imbalanced datasets used in lesion classification studies. The study [49] demonstrated that models with high overall accuracy may still exhibit substantially lower sensitivity for minority lesion classes, indicating that overall accuracy can conceal class-wise performance disparities. Similarly, study [57] showed that a model with good overall accuracy did not achieve comparable sensitivity or minority-class F1-score. Consequently, models reporting near-perfect accuracy cannot be assumed to provide equally reliable detection across all lesion types. AUC was widely adopted as it is less sensitive to class imbalance and provides a threshold-independent measure of discrimination. Nevertheless, AUC evaluates ranking ability rather than actual clinical decision-making performance. A model with a high AUC may still operate at a clinically unsuitable threshold, producing excessive false positives [54] or false negatives [20], [47], [49]. Furthermore, several studies demonstrated performance degradation [20], [9], [54], [55] during external validation despite strong internal AUC values, suggesting that discrimination performance alone is insufficient to establish model robustness or clinical applicability. Sensitivity and specificity introduce an additional challenge because they often exhibit an inverse relationship. Prioritizing sensitivity reduces the likelihood of missed malignant lesions, which is particularly important in cancer diagnosis, but may increase false-positive findings and unnecessary follow-up investigations [9], [51], [54]. Conversely, optimizing specificity reduces false alarms but risks overlooking clinically significant lesions [57]. This trade-off indicates that neither metric should be interpreted independently, and their relative importance depends on the intended clinical application. Precision and F1-score provide complementary information regarding prediction reliability and detection performance. However, precision does not account for missed lesions and may therefore overlook clinically important false negatives [20]. Similarly, although F1-score balances precision and sensitivity, it summarizes performance into a single value and may mask important trade-offs between lesion detection and false-positive findings [50].

It has been noticed that across the reviewed studies, statistical significance testing is inconsistently used. Statistical tests such as DeLong, McNemar, permutation, paired t-tests, and JAFROC analysis strengthen the credibility of performance comparisons by determining whether observed improvements are statistically significant rather than arising from random variation. However, many studies reported higher AUC, accuracy, or F1-score without conducting such analyses. Consequently, the reliability of these reported improvements and the superiority of the proposed models over competing approaches cannot always be established with confidence.

Qualitative evaluations such as radiologist comparisons and visualizations complement quantitative metrics by improving the interpretability of model predictions. Overall, these findings highlight that high numerical performance alone is insufficient for clinical adoption. Comprehensive evaluation should integrate quantitative metrics, statistical validation, and interpretability to provide a more reliable assessment of model performance.

4.4 Shared challenges and transferable strategies

Despite differences in lesion type and anatomical location, the reviewed studies revealed several common challenges across CT-based lesion classification. Substantial variations in lesion appearance, including low contrast, irregular or diffuse boundaries, and overlapping morphology, remained major obstacles

for accurate classification across multiple organs. These challenges were further exacerbated by limited and imbalanced annotated datasets, particularly for malignant or rare lesion subtypes, and by the reliance on single-center datasets. These can increase the risk of overfitting and limit model generalizability. Although recent architectures have achieved excellent internal performance, external validation has been comparatively limited, indicating that robust clinical generalization remains an unresolved challenge. Importantly, the lack of model interpretability further undermines clinical confidence, as strong numerical performance rarely translates into transparent or explainable decision-making.

Alongside these shared challenges, the literature identifies several transferable modeling strategies that repeatedly demonstrate effectiveness across different organs and classification tasks. Transfer learning with pretrained CNN backbones, particularly ResNet-based architectures, consistently improves robustness and generalization in lung and kidney studies [35, 64]. In comparison, VGG-based architectures remain effective for lung and liver lesion classification despite limited data availability [40, 57]. Rather than relying on isolated architectural innovations, successful approaches typically integrate complementary design principles. Hybrid architectures, combining CNNs with attention mechanisms or Transformer-based global reasoning, have been widely adopted in lung, liver, and pancreatic lesion classification to improve contextual understanding and suppress irrelevant background information [51, 52, 54]. Pure Transformer-based models further enhance this capability by capturing long-range spatial dependencies through self-attention, enabling more effective modeling of complex anatomical relationships across organs [41, 58]. Segmentation-guided classification hybrid frameworks have shown consistent benefits in stomach and kidney imaging by constraining the classifier's focus to anatomically relevant regions and reducing background-induced bias [38, 53]. In parallel, bagging-based ensemble strategies improve prediction stability by capturing lesion characteristics across multiple resolutions and reducing sensitivity to noise and anatomical variability, particularly in lung and liver imaging [9, 49]. SSL models learn general features from large unlabeled data that can be reused and fine-tuned for different organs or tasks with limited labeled datasets [32, 43, 59].

Overall, these findings confirm that shared challenges across organs shape CT-based lesion classification and that progress depends on transferable, data-aware modeling strategies rather than organ-specific optimization, underscoring the value of a multi-organ analytical perspective.

4.5 Future directions and research challenges

This review highlights several critical research challenges and opportunities that should be addressed to advance the CT-based lesion classification systems. A major challenge remains the detection of lesions with diffuse margins, low contrast, and overlapping textures, which frequently obscure clear lesion boundaries and lead to misclassification. Improving the modeling of long-range contextual relationships and spatial dependencies, therefore, represents an important research direction. Architectures capable of integrating local morphological cues with broader anatomical context [33, 41, 54] may enable more robust differentiation between subtle lesion patterns and surrounding tissues, particularly in complex imaging scenarios.

Another persistent limitation is the scarcity and imbalance of annotated datasets, especially for rare or malignant lesion subtypes. Many existing models are trained on relatively small datasets, increasing the risk of biased predictions and limited generalizability. Compared to an imbalanced dataset, a balanced dataset DL model provides better performance [43]. Addressing this issue requires greater exploration of synthetic data generation approaches, such as GAN-based augmentation [33, 46, 48, 49, 53], to enrich minority-class samples and improve dataset diversity. Moreover, SSL [32, 43, 59] should be further explored to leverage large volumes of unlabeled CT images, enabling models to learn robust feature representations and reducing dependence on limited annotated data.

Improving model generalizability across diverse clinical environments remains a major research challenge. Although several reviewed studies incorporated external validation using independent institutions, public datasets, or different imaging protocols, external evaluation was often limited to one or a few cohorts and rarely included geographically diverse multi-center populations. Future research should therefore prioritize large-scale multi-center collaborations, standardized external validation protocols, and direct comparison with radiologist performance to establish robust clinical reliability [35], [58]. In addition, collaborative learning paradigms such as federated learning [71] offer a promising direction for developing generalizable models across multiple healthcare institutions while preserving patient privacy and accommodating heterogeneous imaging environments.

Current evaluation practices require greater standardization. Many studies primarily reported improvements in accuracy or AUC without comprehensive statistical significance analysis, external validation, or calibration assessment. Future studies should adopt standardized evaluation protocols incorporating multiple complementary performance metrics, statistical significance testing, calibration analysis, decision curve analysis, and multi-center validation to provide more reliable evidence of clinical effectiveness, as in study [54].

Improving the interpretability of deep learning models remains an important research direction, as limited transparency continues to hinder clinical adoption. Explainable AI techniques, including coarse visualization approaches such as CAM and Grad-CAM [21, 43, 46, 51–54, 58, 59], and feature-attribution techniques such as SHAP and LIME [44], provide valuable insights into model decision processes by highlighting image regions and features influencing predictions. Integrating these techniques into diagnostic frameworks can improve model transparency, support clinician interpretation, and enhance trust in AI-assisted clinical decision-making.

Addressing the computational demands of advanced deep learning models is another significant research direction. Many current architectures require substantial computational resources and long training cycles, which limit their practical deployment in clinical environments. Developing computationally efficient architectures, lightweight models, together with the use of transfer learning, can significantly reduce training time and resource requirements [28, 35, 41, 52]. These strategies will facilitate scalable deployment and enable real-time integration of AI systems into clinical workflows. By addressing these research directions, future studies can overcome current limitations and advance the development of robust, interpretable, and clinically relevant deep learning systems for CT-based lesion classification.

5 Conclusion

This systematic literature review comprehensively investigated recent advances in DL for CT-based lesion classification across multiple organs, including the lung, liver, kidney, pancreas, and stomach, by addressing four research questions related to model architectures, architectural influence, training strategies, and evaluation metrics. The review demonstrates a clear evolution from CNNs toward hybrid, Transformer-based, and SSL frameworks, reflecting the growing demand for models capable of capturing both local lesion characteristics and long-range contextual information.

The analysis indicates that architectural design is a key determinant of classification performance and generalizability. Conventional CNNs remain effective for extracting local texture and morphological features, whereas hybrid and Transformer-based architectures consistently achieve more robust and balanced performance by integrating feature fusion, attention mechanisms, and global contextual modeling. ResNet emerged as the most widely adopted backbone across diverse architectures because of its flexibility and strong feature extraction capability. Transfer learning and advanced data augmentation were identified as the most frequently employed training strategies, substantially improving feature representation, reducing overfitting, and enhancing performance when annotated CT datasets were limited. However, no single architecture or training strategy consistently outperformed all others, with model effectiveness remaining highly dependent on dataset characteristics and lesion complexity.

The review further reveals that many reported performance improvements should be interpreted with caution. Most studies relied on imbalanced, small datasets with limited external validation, while statistical significance testing and explainability analyses were inconsistently performed. Consequently, commonly reported metrics, such as accuracy and AUC, may not fully reflect real-world clinical performance, particularly for minority lesion classes and heterogeneous patient populations. In addition, although LDCT is clinically important, particularly for lung cancer screening, relatively few studies specifically investigated LDCT-based lesion classification, highlighting an important research gap.

Overall, this review provides a comprehensive synthesis of the current state of DL-based CT lesion classification by integrating evidence across multiple organs, model architectures, training strategies, evaluation methodologies, and clinical challenges. The findings suggest that future research should prioritize multi-center and externally validated datasets, standardized evaluation protocols, improved model interpretability, and computationally efficient architectures that balance predictive performance with clinical feasibility. By identifying shared challenges and transferable methodological strategies across different organs, this review provides practical guidance for developing more robust, generalizable, and clinically relevant deep learning systems for CT-based lesion classification.

Declarations

Author Contribution

UMMEY HANY AINAN conducted the literature search, study selection, quality assessment, data extraction, analysis, and drafted the manuscript. NOR ANIZA ABDULLAH, K.A.S.H. KULATHILAKE, AZNUL QALID Md SABRI, JEANNIE WONG, KHIN WEE LAI, SUZAN J. OBAIYS contributed to the review process, interpretation of findings, and critical revision of the manuscript. NOR ANIZA ABDULLAH and K.A.S.H. KULATHILAKE supervised the overall study. All authors reviewed and approved the final manuscript.

Acknowledgement

This work was supported by the Fundamental Research Grant Scheme (FRGS) of the Ministry of Higher Education, Malaysia, Project Number FP061-2024.

References

1. Bizuayehu, H. M., et al. (2024). Global disparities of cancer and its projected burden in 2050. *JAMA network open*, *7*(11), e2443198–e2443198.
2. Ahmad, I., & Alqurashi, F. (2024). Early cancer detection using deep learning and medical imaging: A survey. *Critical Reviews in Oncology/Hematology*, *204*, 104528.
3. Wei, Y., et al. (2024). Focal liver lesion diagnosis with deep learning and multistage CT imaging. *Nature communications*, *15*(1), 7040.
4. Ma, K., Zheng, M., Chen, W., Qi, Y., & Rong, H. Research progress in computer-aided diagnosis systems for lung cancer, *npj Digital Medicine*, vol. 8, no. 1, p. 722, 2025/11/26 2025. 10.1038/s41746-025-02101-3
5. Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R., & Bridge, C. P. (2025). A review of deep learning for brain tumor analysis in MRI. *NPJ Precision Oncology*, *9*(1), 2.
6. Abu Abeelh, E., & AbuAbeileh, Z. (2024). Comparative Effectiveness of Mammography, Ultrasound, and MRI in the Detection of Breast Carcinoma in Dense Breast Tissue: A Systematic Review, (in eng), *Cureus*, vol. 16, no. 4, p. e59054, Apr 10.7759/cureus.59054
7. Naseri, H., & Safaei, A. A. (2025). Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review. *BMC cancer*, *25*(1), 75.
8. Astaraki, M., Zakko, Y., Dasu, I. T., Smedby, Ö., & Wang, C. (2021). Benign-malignant pulmonary nodule classification in low-dose CT with convolutional features. *Physica Medica*, *83*, 146–153.
9. Ahmadyar, Y., Kamali-Asl, A., Samimi, R., Arabi, H., & Zaidi, H. Automated pulmonary nodule classification from low-dose CT images using ERBNet: an ensemble learning approach, *Medical & Biological Engineering & Computing*, vol. 63, no. 9, pp. 2767–2779, 2025/09/01 2025. 10.1007/s11517-025-03358-2
10. Jiang, H., et al. (2023). A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Computers in Biology and Medicine*, *157*, 106726.

11. Survarachakan, S., et al. (2022). Deep learning for image-based liver analysis—A comprehensive review focusing on malignant lesions. *Artificial Intelligence in Medicine*, *130*, 102331.
12. Nayantara, P. V., Kamath, S., Manjunath, K., & Rajagopal, K. (2020). Computer-aided diagnosis of liver lesions using CT images: A systematic review. *Computers in Biology and Medicine*, *127*, 104035.
13. de Margerie-Mellon, C., & Chassagnon, G. (2023). Artificial intelligence: A critical review of applications for lung nodule and lung cancer. *Diagnostic and interventional imaging*, *104*(1), 11–17.
14. Thanoon, M. A., Zulkifley, M. A., Mohd Zainuri, M. A. A., & Abdani, S. R. (2023). A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images, *Diagnostics*, vol. 13, no. 16, p. 2617, [Online]. Available: https://www.mdpi.com/2075-4418/13/16/2617
15. Sachdeva, M., & Kushwaha, A. K. S. (2023). The power of deep learning for intelligent tumor classification systems: A review. *Computers and Electrical Engineering*, *106*, 108586.
16. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, *Bmj*, vol. 339.
17. Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of epidemiology & community health*, *52*(6), 377–384.
18. Rahman, H., Bukht, T. F. N., Imran, A., Tariq, J., Tu, S., & Alzahrani, A. (2022). A deep learning approach for liver and tumor segmentation in CT images using ResUNet, *Bioengineering*, vol. 9, no. 8, p. 368.
19. Zhou, J., et al. (2021). Automatic detection and classification of focal liver lesions based on deep convolutional neural networks: a preliminary study. *Frontiers in oncology*, *10*, 581210.
20. Hu, J. (2025). Hypermetabolic pulmonary lesions detection and diagnosis based on PET/CT imaging and deep learning models. *European Journal of Nuclear Medicine and Molecular Imaging*, pp. 1–15.
21. Zhang, Y., et al. (2023). PB-LNet: a model for predicting pathological subtypes of pulmonary nodules on CT images. *BMC cancer*, *23*(1), 936.
22. Marentakis, P., et al. (2021). Lung cancer histology classification from CT images based on radiomics and deep learning models. *Medical & biological engineering & computing*, *59*(1), 215–226.
23. Zeng, H. (2025). MSSFN: A multi-scale sequence fusion network for CT-based diagnosis of pulmonary complications, *Neurocomputing*, vol. 634, p. 129878.
24. Giandola, T. (2023). Imaging in gastric cancer: current practice and future perspectives, *Diagnostics*, vol. 13, no. 7, p. 1276.
25. Jabłońska, B., Gudz, A., Hinborch, T., Bujała, B., Biskup, K., & Mrowiec, S. (2023). Pancreatic cystic tumors: a single-center observational study, *Medicina*, vol. 59, no. 2, p. 241.
26. Bazzocchi, M. V. (2023). The role of CT imaging in characterization of small renal masses, *Diagnostics*, vol. 13, no. 3, p. 334.
27. Armato, S. G. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, *38*(2), 915–931.
28. Saied, M., Raafat, M., Yehia, S., & Khalil, M. M. (2023). Efficient pulmonary nodules classification using radiomics and different artificial intelligence strategies. *Insights into Imaging*, *14*(1), 91.
29. Katase, S., et al. (2022). Development and performance evaluation of a deep learning lung nodule detection system. *BMC Medical Imaging*, *22*(1), 203.
30. Shanshool, A. M., Bouchakwa, M., & Amor, I. A. B. (2023). Improved accuracy of pulmonary nodule classification on LIDC-IDRI dataset using deep learning. *Procedia Computer Science*, *225*, 394–403.
31. Jena, S. R., & George, S. T. (2020). Morphological feature extraction and KNG-CNN classification of CT images for early lung cancer detection. *International Journal of Imaging Systems and Technology*, *30*(4), 1324–1336.
32. Huang, H., Wu, R., Li, Y., & Peng, C. (2022). Self-supervised transfer learning based on domain adaptation for benign-malignant lung nodule classification on thoracic CT. *IEEE Journal of Biomedical and Health Informatics*, *26*(8), 3860–3871.
33. Jin, H., Yu, C., Zhang, J., Zheng, R., Fu, Y., & Zhao, Y. (2025). Multitask Swin Transformer for classification and characterization of pulmonary nodules in CT images. *Quantitative Imaging in Medicine and Surgery*, *15*(3), 1845–1861.
34. Wang, N. *Luna 16*, 10.21227/0kjp-g187
35. Cui, S., et al. (2020). Development and clinical application of deep learning model for lung nodules screening on CT images. *Scientific reports*, *10*(1), 13657.
36. Armato, S. G. (2015). III et al., SPIE-AAPM-NCI lung nodule classification challenge dataset. *(No Title)*.
37. Venkadesh, K. V. Deep Learning for Malignancy Risk Estimation of Pulmonary Nodules Detected at Low-Dose Screening CT, *Radiology*, vol. 300, no. 2, pp. 438–447, 2021/08/01 2021. 10.1148/radiol.2021204433
38. Liu, J., Yildirim, O., Akin, O., & Tian, Y. (Jan 13 2023). AI-Driven Robust Kidney and Renal Mass Segmentation and Classification on 3D CT Images. *Bioengineering (Basel)*, *10*(1). 10.3390/bioengineering10010116
39. Alyasriy, H. A. H. Muayed. *The IQ-OTH/NCCD lung cancer dataset*, 10.17632/bhmdr45bh2.4
40. Humayun, M., Sujatha, R., Almuayqil, S. N., & Jhanjhi, N. (2022). A transfer learning approach with a convolutional neural network for the classification of lung carcinoma. *Healthcare* (Vol. 10, p. 1058). MDPI. 6.
41. Martín, Ó. A., & Sánchez, J. (2025). Evaluation of Vision Transformers for Multi-Organ Tumor Classification Using MRI and CT Imaging, *Electronics*, vol. 14, no. 15, p. 2976.

42. Islam, M. N., Hasan, M., Hossain, M. K., Alam, M. G. R., Uddin, M. Z., & Soylu, A. (Jul 6 2022). Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *(in eng) Sci Rep*, *12*(1), 11440. 10.1038/s41598-022-15634-4
43. Özbay, E., Özbay, F. A., & Gharehchopogh, F. S. (2024). Kidney tumor classification on CT images using self-supervised learning. *Computers in Biology and Medicine*, *176*, 108554.
44. Bhandari, M., Yogarajah, P., Kavitha, M. S., & Condell, J. (2023). Exploring the capabilities of a lightweight CNN model in accurately identifying renal abnormalities: Cysts, stones, and tumors, using LIME and SHAP. *Applied Sciences*, *13*(5), 3125.
45. Alzu'bi, D. (2022). Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans, (in eng), *J Healthc Eng*, vol. p. 3861161, 2022. 10.1155/2022/3861161
46. Gupta, P., et al. (Jan 2025). Automatic localization and deep convolutional generative adversarial network-based classification of focal liver lesions in computed tomography images: A preliminary study. *Journal Of Gastroenterology And Hepatology*, *40*(1), 166–176. 10.1111/jgh.16803
47. Yu, J., Yang, B., Wang, J., Leader, J., Wilson, D., & Pu, J. (2020). 2D CNN versus 3D CNN for false-positive reduction in lung cancer screening. *Journal of Medical Imaging*, *7*(5), 051202–051202.
48. Cay, T. (2025). Lung cancer diagnosis with GAN supported deep learning models. *Bio-Medical Materials and Engineering*, p. 09592989241308775.
49. Lee, H., Lee, H., Hong, H., Bae, H., Lim, J. S., & Kim, J. (2021). Classification of focal liver lesions in CT images using convolutional neural networks with lesion information augmented patches and synthetic data augmentation. *Medical Physics*, *48*(9), 5029–5046. https://doi.org/10.1002/mp.15118
50. Jafery, N. N., Sulaiman, S. N., Osman, M. K., Abdul Karim, N. K., & Che Soh, Z. H. (2025). and N. A. Mat Isa, Comparative Analysis of Hybrid 1D-CNN-LSTM and VGG16-1D-LSTM for Lung Lesion Classification. *Journal of Electrical Engineering & Technology*, pp. 1–14.
51. Zhang, Y. (2023). Deep-learning model of ResNet combined with CBAM for malignant–benign pulmonary nodules classification on computed tomography images, *Medicina*, vol. 59, no. 6, p. 1088.
52. Lou, M., Ying, H., Liu, X., Zhou, H. Y., Zhang, Y., & Yu, Y. (2025). SDR-Former: A Siamese Dual-Resolution Transformer for liver lesion classification using 3D multi-phase imaging, *Neural Netw*, vol. 185, p. 107228, May 10.1016/j.neunet.2025.107228
53. Gao, Z., et al. (2023). Development of a deep learning model for early gastric cancer diagnosis using preoperative computed tomography images. *Frontiers In Oncology*, *13*, 1265366. 10.3389/fonc.2023.1265366
54. Liao, H., et al. (Apr 2025). Deep learning-based MVIT-MLKA model for accurate classification of pancreatic lesions: a multicenter retrospective cohort study. *La Radiologia Medica*, *130*(4), 508–523. 10.1007/s11547-025-01949-5
55. Fu, C. L., et al. (2023). Discrimination of ground-glass nodular lung adenocarcinoma pathological subtypes via transfer learning: A multicenter study. *Cancer Medicine*, *12*(18), 18460–18469.
56. Gai, L., Xing, M., Chen, W., Zhang, Y., & Qiao, X. (2024). Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective? *Multimedia Tools and Applications*, *83*(20), 59253–59269.
57. Lee, H., Kim, D., Lim, J., & Hong, H. (Oct 2025). Class-Wise Combination of Mixture-Based Data Augmentation for Class Imbalance Learning of Focal Liver Lesions in Abdominal CT Images. *J Imaging Inform Med*, *38*(5), 2983–2993. 10.1007/s10278-025-01415-8
58. Wen, Y., et al. (2025). Predicting pathological subtypes of pure ground-glass nodules using Swin Transformer deep learning model. *Insights Into Imaging*, *16*(1), 223.
59. Song, J., Hu, Y., Wang, H., & Chen, Y. W. (2025). Liver-VLM: Enhancing Focal Liver Lesion Classification with Self-Supervised Vision-Language Pretraining, Applied Sciences, 15, 23, p. 12578.
60. Lin, C. Y., et al. (2024). Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT. *La radiologia medica*, *129*(1), 56–69.
61. UrRehman, Z., et al. (2024). Effective lung nodule detection using deep CNN with dual attention mechanisms. *Scientific Reports*, *14*(1), 3934.
62. Zhang, M., Sun, L., Li, X., Zhu, W., Yi, Y., & Yan, F. (2025). A multi-scale self-calibrating lung nodule detection based on SPC-UNet. *Biomedical Signal Processing and Control*, *104*, 107515.
63. Ahmadyar, Y., Kamali-Asl, A., Arabi, H., Samimi, R., & Zaidi, H. (2024). Hierarchical approach for pulmonary-nodule identification from CT images using YOLO model and a 3D neural network classifier. *Radiological physics and technology*, *17*(1), 124–134.
64. Zhao, T., Sun, Z., Guo, Y., Sun, Y., Zhang, Y., & Wang, X. (2023). Automatic renal mass segmentation and classification on CT images based on 3D U-Net and ResNet algorithms, (in eng). *Frontiers In Oncology*, *13*, 1169922. 10.3389/fonc.2023.1169922
65. Pimpalkar, A., Saini, D. K. J. B., Shelke, N., Balodi, A., Rapate, G., & Tolani, M. (2025). Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Scientific Reports*, *15*(1), 10741.
66. Tanaka, T., et al. (2020). Differentiation of small (≤ 4 cm) renal masses on multiphase contrast-enhanced CT by deep learning. *American journal of roentgenology*, *214*(3), 605–612.
67. Uhm, K. H., et al. (2021). Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *NPJ precision oncology*, *5*(1), 54.
68. Astaraki, M., Yang, G., Zakko, Y., Toma-Dasu, I., Smedby, O., & Wang, C. (2021). A comparative study of radiomics and deep-learning based methods for pulmonary nodule malignancy prediction in low dose CT images. *Frontiers in oncology*, *11*, 737368.
69. Gedeon, K. K., & Liu, Z. (2024). Classification of liver lesions in CT images based on LivlesioNet, modified Multi-Scale CNN with bridge Scale method. *Multimedia Tools and Applications*, *83*(3), 8911–8929.
70. Yang, P., Wu, J., Liu, M., Zheng, Y., Zhao, X., & Mao, Y. (2024). Preoperative CT-based radiomics and deep learning model for predicting risk stratification of gastric gastrointestinal stromal tumors. *Medical Physics*, *51*(10), 7257–7268. https://doi.org/10.1002/mp.17276

71. Al-Saleh, A., Tejani, G. G., Mishra, S., Sharma, S. K., & Mousavirad, S. J. (2025). A federated learning-based privacy-preserving image processing framework for brain tumor detection from CT scans. *Scientific Reports*, *15*(1), 23578.
72. Abu-Azzam, O. H. (2025). Hybrid ResNet-ViT Framework for Endometrial Lesion Analysis: A Comparative Study of MRI and CT in Endometrial Cancer Classification. *International Journal of Women's Health* pp. 4103–4130.
73. Zhu, S. (2025). STD-Net: a spatio-temporal decoupling network for multiphasic liver lesion segmentation and characterization. *npj Digital Medicine*.
74. Yuan, M. (2023). Cluster-induced mask transformers for effective opportunistic gastric cancer screening on non-contrast CT scans, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, : Springer, pp. 146–156.
75. Khaliq, A., et al. (2025). Revolutionizing medical imaging: A cutting-edge AI framework with vision transformers and perceiver IO for multi-disease diagnosis. *Computational Biology and Chemistry*, *119*, 108586.
76. Sugawara, K., et al. (2025). Breast cancer classification based on breast tissue structures using the Jigsaw puzzle task in self-supervised learning. *Radiological Physics and Technology*, *18*(1), 209–218.
77. Fan, Z. (2024). Self-supervised learning based on StyleGAN for medical image classification on small labeled dataset. *Medical Imaging 2024: Image Processing* (Vol. 12926, pp. 727–733). SPIE.

Figures

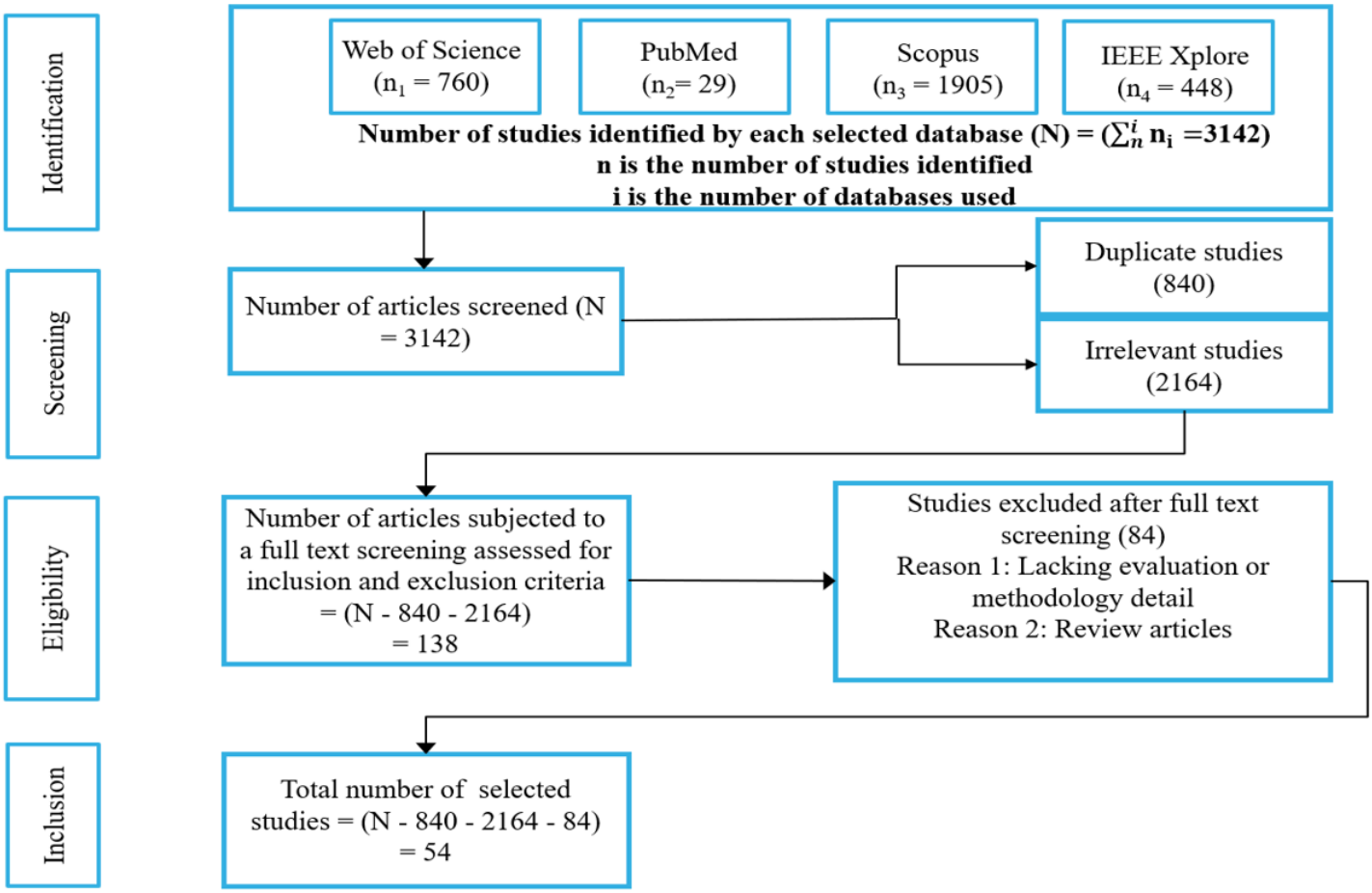


Figure 1

PRISMA flowchart for the study selection process

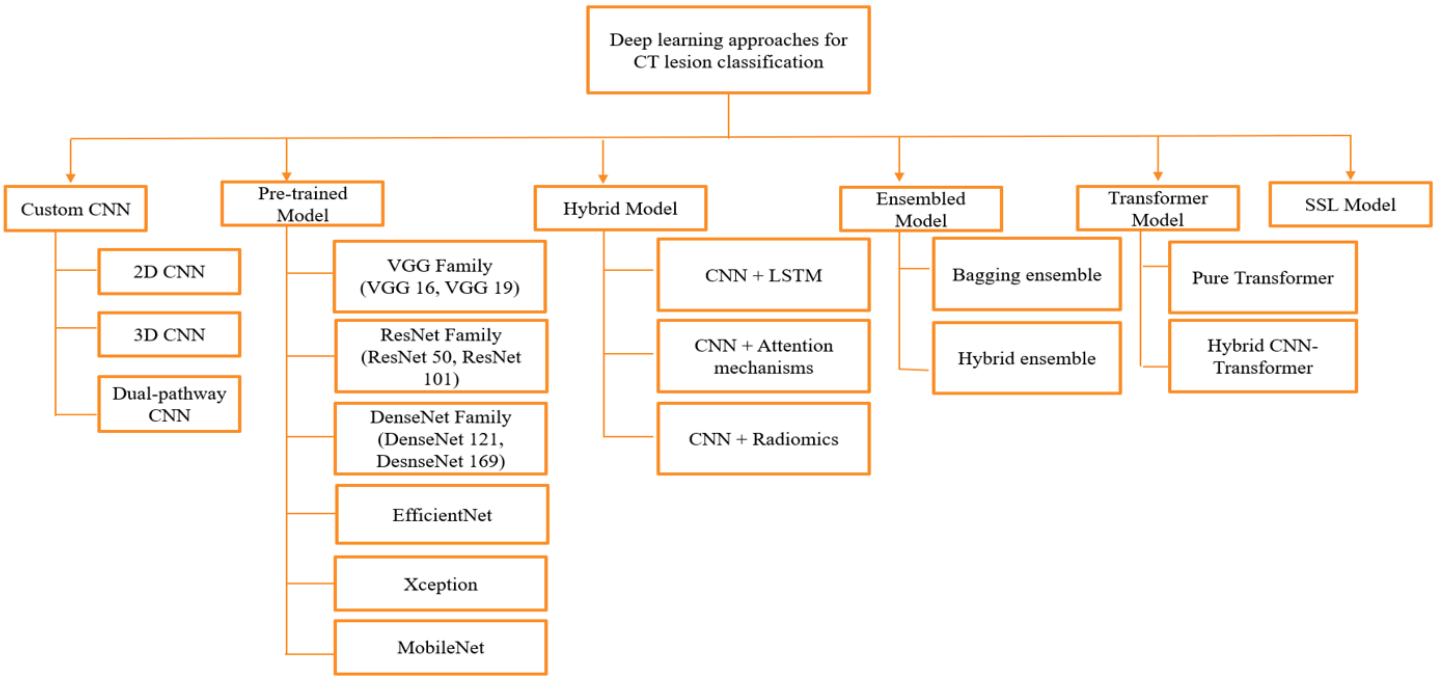


Figure 2

Taxonomy of DL approaches for CT lesion classification

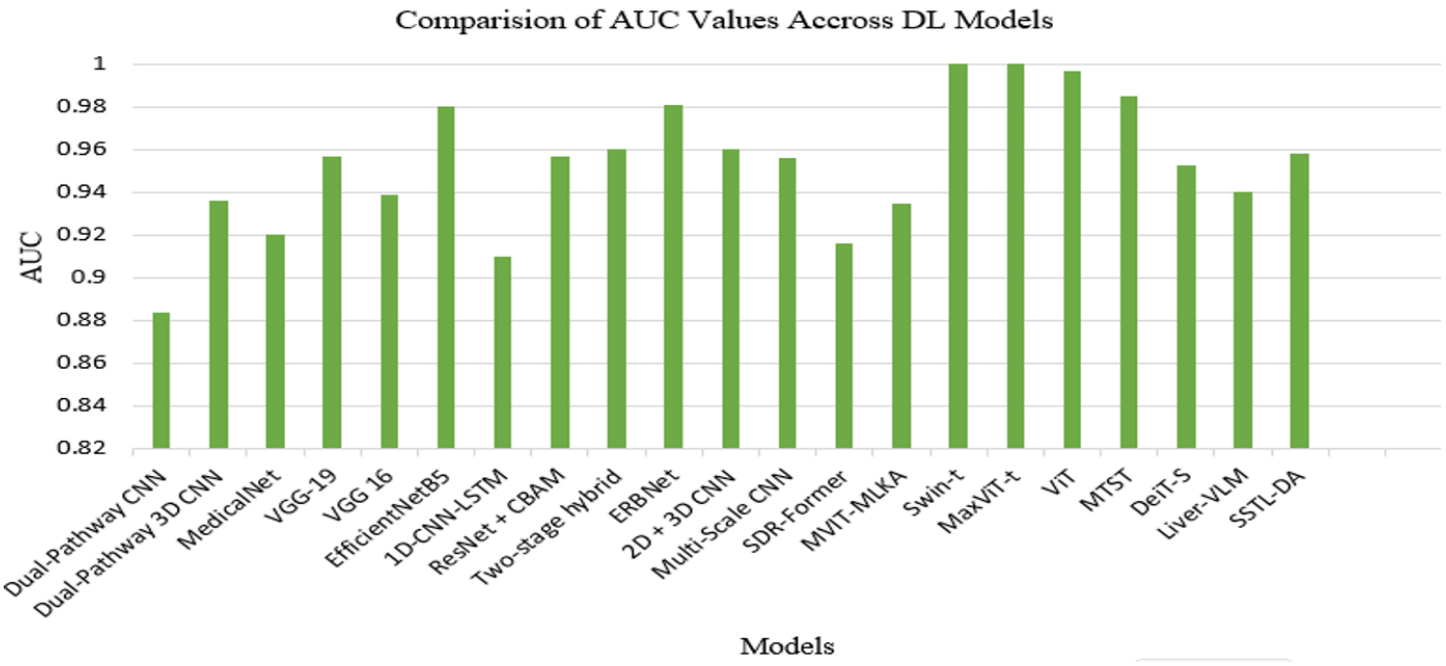


Figure 3

Comparative AUC performance of different DL models for CT lesion classification

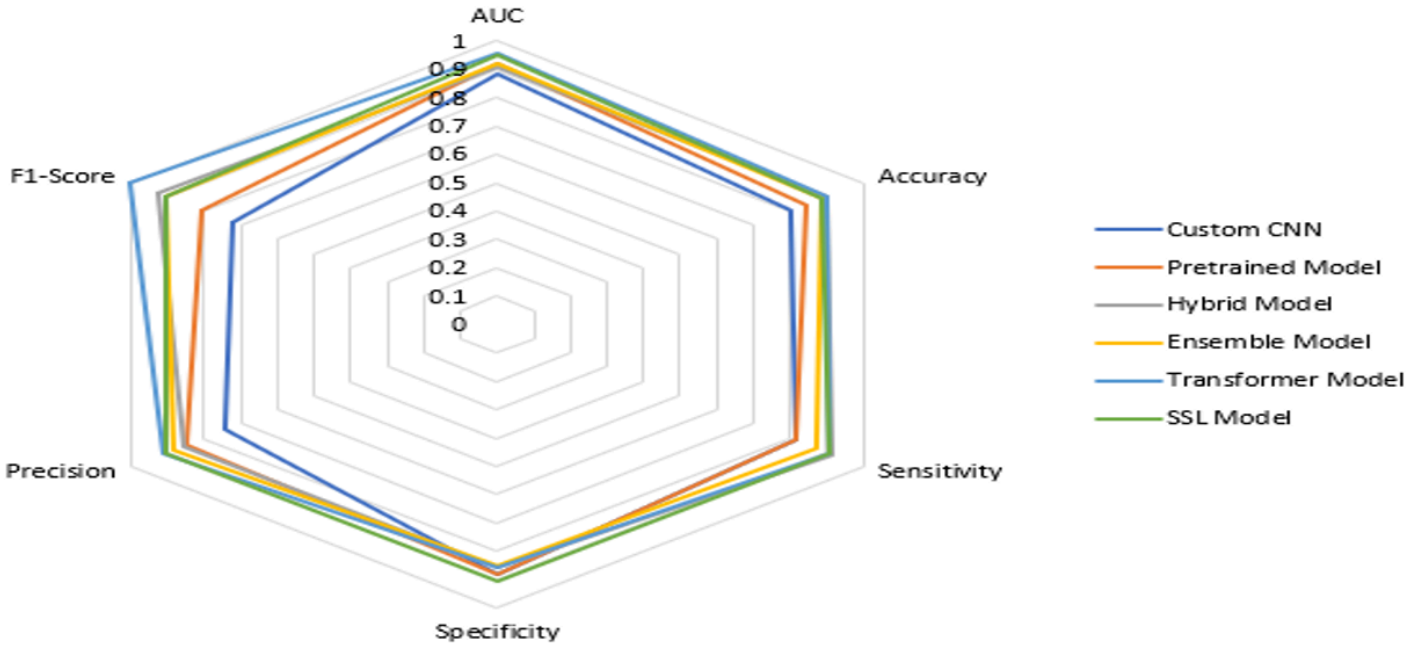


Figure 4

Radar chart comparing the average performance of different DL model categories across evaluation metrics