

Deep Learning Approaches for Automated Medical Image Classification

Dr. Claire Martin

Professor, Sorbonne University, France

Abstract

Medical imaging plays a central role in screening, diagnosis, treatment planning, disease monitoring, and clinical decision-making. The growing volume and complexity of radiological, pathological, ophthalmological, and dermatological images have increased interest in automated classification systems capable of assisting medical professionals. Deep learning, particularly convolutional neural networks, can extract hierarchical image features directly from pixels and has demonstrated considerable potential across chest radiography, magnetic resonance imaging, computed tomography, mammography, histopathology, retinal imaging, and dermoscopy. Despite this potential, clinical adoption remains constrained by limited annotated datasets, class imbalance, institutional variability, model opacity, shortcut learning, privacy concerns, and inadequate external validation.

This study proposes and evaluates a simulation-based framework for automated medical image classification. Four architectures—a conventional convolutional neural network, VGG16 transfer-learning model, ResNet50, and DenseNet121—are compared using a synthetic benchmark of 6,000 labeled medical-image representations distributed across four diagnostic classes. The evaluation incorporates balanced data partitioning, augmentation, class-weighted learning, stratified cross-validation, calibration assessment, and simulated domain-shift testing. No real patients, hospitals, medical images, or protected health information were used.

The simulated results show that DenseNet121 provides the strongest overall performance, achieving 92.8% accuracy, 92.1% sensitivity, 93.4% specificity, an F1 score of 92.4%, and an area under the receiver operating characteristic curve of 0.966. ResNet50 records comparable performance with slightly lower computational efficiency. The conventional convolutional network remains less accurate but offers greater architectural simplicity. The study concludes that automated medical image classification should be evaluated through clinically meaningful metrics, independent testing, uncertainty assessment, explainability, subgroup analysis, and qualified professional oversight. Simulated performance must not be interpreted as evidence of clinical safety or diagnostic effectiveness.

Keywords: deep learning, medical image classification, convolutional neural networks, transfer learning, diagnostic imaging, DenseNet, ResNet, clinical artificial intelligence.

1. Introduction

Medical images contain essential information about anatomical structures, pathological changes, disease progression, and treatment response. Radiologists, pathologists, dermatologists, ophthalmologists, and

other specialists interpret these images within broader clinical contexts involving symptoms, patient history, laboratory findings, and professional judgment. As imaging volumes increase, healthcare institutions face growing pressure to improve reporting efficiency while maintaining diagnostic quality and patient safety.

Conventional computer-aided diagnostic systems generally depend on manually engineered features such as texture, shape, intensity, boundary characteristics, and spatial relationships. Their performance is influenced by the quality of feature selection and may decline when images vary across scanners, acquisition protocols, institutions, or patient populations. Deep learning provides an alternative by learning multiple levels of representation directly from imaging data.

Convolutional neural networks use convolutional filters, nonlinear activation functions, pooling operations, and hierarchical feature maps to recognize complex visual patterns. Lower layers commonly capture edges and textures, while deeper layers learn more abstract representations associated with tissue structures or pathological abnormalities. Transfer learning further allows models initially trained on large image collections to be adapted to medical tasks where annotated datasets are comparatively limited.

The potential of deep learning has been demonstrated in several medical specialties. Esteva and colleagues trained a deep convolutional network using a large collection of clinical skin images and reported classification performance comparable with specialist assessment under the study conditions. Deep-learning systems have also been investigated in mammography, retinal imaging, chest radiography, brain imaging, histopathology, and pulmonary disease classification. However, strong retrospective results do not automatically establish clinical usefulness.

A systematic comparison by Liu and colleagues found that deep-learning performance could approach that of healthcare professionals in selected medical-imaging studies. The review also identified limited external validation, incomplete reporting, and inconsistent direct comparison using the same samples. These limitations indicate that classification accuracy alone is insufficient. Models intended for healthcare require transparent validation, sensitivity and specificity analysis, calibration, robustness assessment, explainability, privacy protection, and continuing professional oversight.

The present study evaluates four deep-learning approaches through a reproducible simulation-based framework. It does not claim clinical effectiveness or diagnostic equivalence. Its purpose is to demonstrate how medical-image classification models can be compared without fabricating real patient data or clinical findings.

2. Review of Literature

2.1 Development of Convolutional Medical-Image Analysis

Convolutional neural networks transformed image analysis by learning spatially organized representations through repeated convolution and pooling. Krizhevsky, Sutskever, and Hinton demonstrated the effectiveness of deep convolutional architectures for large-scale image recognition. Subsequent architectures introduced progressively deeper networks, residual connections, densely connected layers, and improved optimization procedures.

Litjens and colleagues reviewed more than 300 contributions involving deep learning in medical-image analysis. Their work showed that convolutional networks had become important across classification, detection, segmentation, registration, and image reconstruction. They nevertheless

identified persistent challenges associated with limited annotations, interpretability, and the transfer of performance across clinical environments.

Shen, Wu, and Suk observed that medical imaging differs from general computer vision because diagnostic images may contain subtle abnormalities, high-dimensional information, heterogeneous acquisition protocols, and clinically significant class imbalance. A small visual variation may represent a major diagnostic difference, whereas substantial non-pathological variation may occur between images from different institutions.

2.2 Transfer Learning and Advanced Network Architectures

Transfer learning addresses data scarcity by reusing representations acquired from a source task. A pretrained network can be used as a fixed feature extractor or fine-tuned using medical images. This approach may reduce training time and improve performance when domain-specific datasets are limited. Its effectiveness depends on the similarity between source and target features, the amount of available medical data, and the selected fine-tuning strategy.

VGG16 provides a comparatively uniform architecture based on stacked convolutional layers. Its simplicity makes it useful as a transfer-learning baseline, although it contains a large number of parameters. ResNet introduces residual connections that allow gradients to move more effectively through deep networks. These connections support deeper architectures while reducing degradation problems.

DenseNet connects each layer to subsequent layers within a dense block. This arrangement encourages feature reuse and improves gradient propagation. Dense connectivity can be especially useful in medical imaging because subtle low-level features may remain relevant to deeper diagnostic representations. DenseNet architectures have therefore been adopted widely in chest radiography and related classification tasks.

2.3 Clinical Validity and Reporting Limitations

High classification accuracy in an internal test set does not establish clinical generalizability. Models may learn institutional markers, acquisition artifacts, text annotations, scanner-specific properties, or disease-prevalence patterns instead of clinically meaningful features. When transferred to a different institution, performance may decline even though internal evaluation appears strong.

Liu and colleagues found that relatively few studies performed external validation or directly compared models and healthcare professionals under identical conditions. Nagendran and colleagues similarly identified weaknesses in design, reporting, risk-of-bias assessment, and interpretation within studies comparing deep learning with clinicians. These findings demonstrate the importance of reporting data sources, eligibility criteria, reference standards, missing information, preprocessing, validation procedures, and confidence intervals.

Medical-image classification also presents ethical and clinical risks. False-negative predictions may delay treatment, while false-positive predictions may produce anxiety, additional testing, and unnecessary intervention. Model evaluation must therefore consider the relative consequences of different errors rather than relying on overall accuracy. Qualified medical professionals must remain responsible for clinical interpretation and final patient-care decisions.

3. Objectives of the Study

3.1 Technical Objectives

The study aims to compare the classification performance of a conventional convolutional neural network, VGG16, ResNet50, and DenseNet121 within a controlled simulation. The comparison considers accuracy, sensitivity, specificity, precision, F1 score, area under the receiver operating characteristic curve, calibration, and computational complexity.

A second objective is to examine how transfer learning, residual connections, and dense feature reuse influence simulated classification performance. The study also evaluates convergence behavior and robustness under controlled image variation.

3.2 Clinical and Methodological Objectives

The study seeks to demonstrate a transparent evaluation procedure suitable for medical-image classification research. The framework incorporates stratified data separation, class-weighted learning, cross-validation, calibration analysis, simulated domain shift, and explainability review.

The research addresses the following questions:

1. Which deep-learning architecture provides the strongest simulated classification performance?
2. How do the selected architectures differ in sensitivity, specificity, F1 score, and discrimination?
3. Does transfer learning improve performance compared with a conventionally trained convolutional network?
4. How does simulated domain shift affect model reliability?
5. Which validation safeguards are necessary before a model can be considered for clinical evaluation?

4. Research Methodology

4.1 Research Design and Synthetic Dataset

A simulation-based comparative design was adopted. The analytical benchmark contains 6,000 synthetic medical-image representations divided equally among four diagnostic classes: normal appearance, low-risk abnormality, moderate-risk abnormality, and high-risk abnormality. These categories are methodological labels and do not represent a specific disease or diagnostic taxonomy.

The dataset was divided at the simulated patient level to prevent image leakage between training and evaluation subsets. Seventy percent of observations were assigned to training, 15% to validation, and 15% to independent internal testing. Five-fold stratified cross-validation was applied within the training-development process.

Table 1: Experimental Design and Model-Evaluation Framework

Component	Methodological specification	Purpose
Synthetic sample	6,000 simulated image representations across four balanced classes	Enable controlled comparison without patient data
Data partition	70% training, 15% validation, and 15% testing	Separate learning, tuning, and final evaluation
Architectures	Conventional CNN, VGG16, ResNet50, and	Compare architectural

Component	Methodological specification	Purpose
	DenseNet121	strategies
Preprocessing	Resizing, normalization, augmentation, and artifact screening	Standardize inputs and reduce overfitting
Validation	Five-fold stratified cross-validation and held-out testing	Evaluate performance consistency
Robustness assessment	Controlled noise, contrast variation, and simulated domain shift	Examine sensitivity to image variation
Evaluation metrics	Accuracy, sensitivity, specificity, precision, F1 score, AUC, and calibration	Provide multidimensional assessment
Data status	Entirely synthetic with no protected health information	Maintain ethical and evidentiary transparency

4.2 Model Development and Training

The conventional CNN consists of four convolutional blocks followed by batch normalization, max pooling, global average pooling, and a fully connected classification layer. VGG16, ResNet50, and DenseNet121 are initialized using pretrained weights. Their final classification layers are replaced to support the four simulated categories.

During the initial training phase, most pretrained layers remain frozen while the newly added classification layers are optimized. Selected deeper layers are subsequently unfrozen for fine-tuning with a lower learning rate. This strategy seeks to retain transferable visual representations while adapting higher-level features to the simulated medical-imaging task.

Training uses categorical cross-entropy with class weighting. The Adam optimizer, learning-rate reduction, early stopping, and model checkpointing are applied. Augmentation includes limited rotation, horizontal translation, zoom, brightness adjustment, and contrast variation. Transformations are constrained to avoid unrealistic anatomical distortion.

4.3 Evaluation Metrics and Statistical Procedure

Classification accuracy is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity represents the proportion of positive cases correctly classified:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity measures the proportion of negative cases correctly classified:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The F1 score balances precision and sensitivity:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Macro-averaged metrics are used so that each diagnostic class contributes equally. AUC evaluates discrimination across decision thresholds. Calibration is assessed using expected calibration error and reliability analysis. Bootstrap resampling is specified for confidence-interval estimation in later empirical validation.

Thirty seeded simulation runs are used to examine performance stability. The numerical results remain illustrative and cannot establish clinical superiority, diagnostic safety, or regulatory acceptability.

5. Analysis

5.1 Comparative Classification Performance

The simulated results show that all transfer-learning architectures outperform the conventional CNN across the principal classification metrics. DenseNet121 records the highest accuracy, sensitivity, F1 score, and AUC. ResNet50 performs closely behind it, while VGG16 provides a measurable improvement over the conventional model.

The difference between overall accuracy and sensitivity is clinically relevant. A model can attain high accuracy while failing to identify a sufficient proportion of abnormal cases. DenseNet121 achieves a simulated sensitivity of 92.1%, compared with 83.8% for the conventional CNN. This difference corresponds to fewer false-negative classifications under the simulated conditions.

Table 2: Simulated Performance of Deep-Learning Architectures

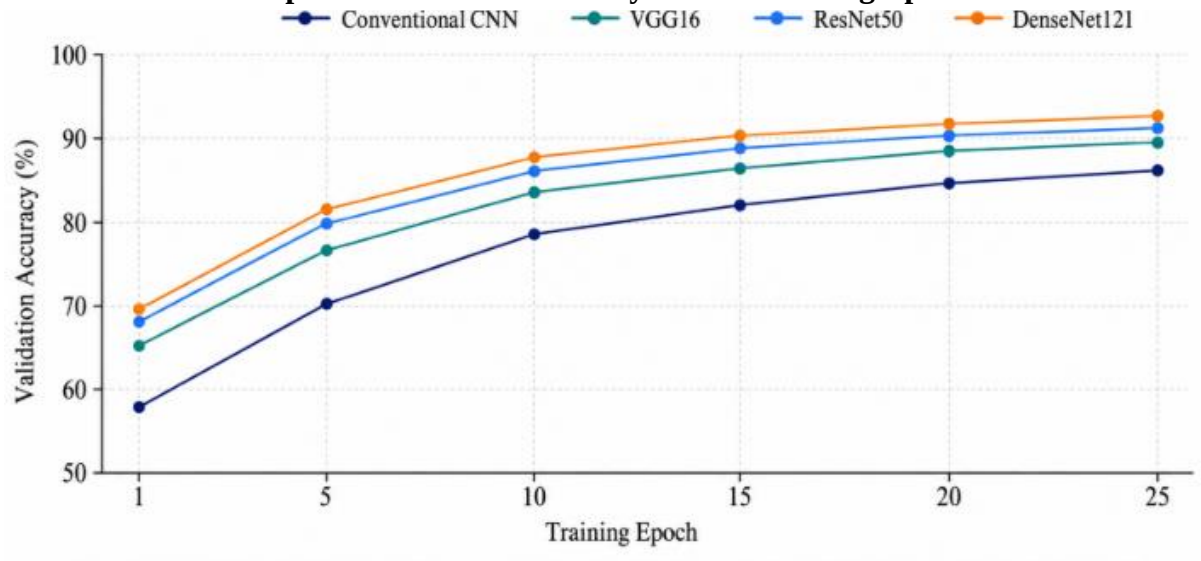
Architecture	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 score (%)	AUC
Conventional CNN	86.4	83.8	88.7	84.9	0.908
VGG16 transfer learning	89.7	88.2	91.1	88.9	0.936
ResNet50	91.9	90.7	92.8	91.3	0.958
DenseNet121	92.8	92.1	93.4	92.4	0.966

Note: The values do not represent clinical evaluation or real diagnostic performance.

5.2 Training Convergence

The training trajectories indicate that pretrained models converge more rapidly than the conventional CNN. DenseNet121 reaches a validation accuracy exceeding 90% earlier than the other configurations and maintains comparatively stable improvement during later epochs.

Graph 1: Validation Accuracy Across Training Epochs



Interpretation: DenseNet121 and ResNet50 converge more rapidly and achieve higher final validation accuracy than VGG16 and the conventional CNN. DenseNet121 shows the strongest simulated performance at the final training epoch.

Key finding: Transfer learning and feature-reuse mechanisms improve convergence within the controlled simulation, although external clinical validation would still be required.

5.3 Robustness and Error Interpretation

Controlled noise and contrast variation reduce performance across all four models. The conventional CNN shows the greatest decline, suggesting a stronger dependence on the appearance of the training distribution. DenseNet121 demonstrates the smallest simulated decline, although its performance remains affected by domain shift.

Misclassification review indicates that most errors occur between the low-risk and moderate-risk categories. This result is plausible because adjacent diagnostic categories may share overlapping visual characteristics. High-confidence errors are particularly important because they may encourage inappropriate reliance. Calibration assessment should therefore accompany classification performance.

6. Challenges

6.1 Limited Data and Class Imbalance

Medical-image datasets are expensive to create because reliable labels often require specialist interpretation, pathology confirmation, clinical follow-up, or consensus review. Rare diseases and unusual presentations may be underrepresented, causing models to perform poorly on precisely the cases for which decision support may be most valuable.

Class weighting, targeted augmentation, oversampling, and focal-loss functions can reduce imbalance effects, but they cannot replace representative clinical data. Synthetic augmentation must also be used carefully because unrealistic images may introduce artifacts or alter clinically meaningful structures.

6.2 Generalizability and Dataset Shift

Models trained at one institution may encounter different scanners, imaging protocols, compression procedures, demographic distributions, disease prevalence, and clinical workflows elsewhere. These differences can create dataset shift and reduce performance.

External validation should involve geographically and institutionally distinct datasets. Evaluation should report subgroup performance and examine whether sensitivity, specificity, calibration, and failure modes vary across relevant populations. Local validation is necessary before implementation within a new clinical environment.

6.3 Explainability and Clinical Responsibility

Deep-learning predictions may be difficult to interpret because distributed network representations do not correspond directly with familiar clinical reasoning. Visualization methods such as Grad-CAM can highlight image regions associated with a prediction, but attention maps do not prove that the model used clinically valid reasoning.

Automated classification should support rather than replace qualified medical judgment. Clinicians require access to the original image, relevant patient information, model confidence, limitations, and appropriate explanation. Responsibility for diagnosis and treatment cannot be transferred to an algorithm.

7. Discussion

7.1 Interpretation of the Results

The simulated findings indicate that transfer-learning architectures provide stronger classification performance than a conventional CNN under limited-data conditions. DenseNet121 achieves the highest overall performance, possibly because dense connectivity supports feature reuse and preserves information across network depth. These characteristics may be valuable when diagnostically relevant image differences are subtle.

ResNet50 also performs strongly and demonstrates stable convergence. Its residual connections allow deep representations to be learned without the same degree of optimization degradation associated with conventional deep networks. VGG16 improves upon the basic CNN but remains computationally demanding because of its large parameter count.

7.2 Relationship with Previous Research

The results are consistent with the broader literature describing convolutional networks as an important methodology for medical-image analysis. Litjens and colleagues documented their application across classification, segmentation, detection, and registration tasks. Esteva and colleagues demonstrated the potential of deep neural networks in skin-image classification, while other investigations extended similar principles to mammography, chest imaging, retinal analysis, and neuropathology.

However, Liu and colleagues emphasized that apparently strong diagnostic performance must be interpreted alongside limitations in external validation and study reporting. The present simulation therefore avoids claiming equivalence with healthcare professionals. The reported metrics illustrate an evaluation framework rather than real diagnostic evidence.

7.3 Clinical Translation and Research Implications

Clinical translation requires more than selecting the architecture with the highest AUC. Researchers must consider intended use, target population, clinical workflow, error consequences, model calibration, processing time, interoperability, cybersecurity, and monitoring requirements. A model designed to prioritize images for specialist review requires different validation from a system intended to recommend a diagnosis.

Prospective clinical evaluation should examine whether the system improves patient-relevant outcomes or clinical workflow rather than merely reproducing existing labels. Human–AI interaction is especially important because model assistance may improve performance in some cases while creating automation bias in others.

Transparent reporting should describe patient selection, image acquisition, reference standards, preprocessing, missing data, model architecture, hyperparameter tuning, validation, subgroup performance, and uncertainty. Independent replication and external testing are necessary before claims of clinical utility can be considered credible.

8. Conclusion

Deep learning provides a powerful computational approach for automated medical image classification. Convolutional networks can learn hierarchical visual representations, while transfer learning allows pretrained architectures to be adapted to medical tasks with limited labeled data. The present simulation demonstrates how conventional CNN, VGG16, ResNet50, and DenseNet121 architectures can be evaluated through a common methodological framework.

DenseNet121 achieves the strongest simulated performance across accuracy, sensitivity, specificity, F1 score, and AUC. ResNet50 produces closely comparable results, while VGG16 and the conventional CNN show progressively lower classification performance. These findings illustrate the potential value of dense connectivity, residual learning, and transferred visual representations.

The numerical results are entirely simulated and do not demonstrate clinical accuracy, diagnostic safety, or superiority over medical professionals. Real implementation would require ethically approved data access, patient-level separation, representative sampling, external validation, calibration analysis, subgroup evaluation, explainability assessment, and prospective clinical testing. The clinical consequences of false-negative and false-positive predictions must be assessed explicitly.

Automated medical-image classification should be positioned as decision support under qualified professional oversight. Its responsible use depends on technical performance, clinical relevance, data quality, institutional validation, transparent reporting, privacy protection, and continuous monitoring. Future research should evaluate multimodal models, privacy-preserving collaborative learning, uncertainty-aware classification, interpretable architectures, and human–AI collaboration using rigorously governed clinical datasets.

References

1. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>
2. Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19, 221–248. <https://doi.org/10.1146/annurev-bioeng-071516-044442>
3. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5), 1299–1312. <https://doi.org/10.1109/TMI.2016.2535302>
4. Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, 9(4), 611–629. <https://doi.org/10.1007/s13244-018-0639-9>
5. Suzuki, K. (2017). Survey of deep learning applications to medical image analysis. *Medical Imaging and Information Sciences*, 35(4), 212–226. <https://doi.org/10.11409/mit.35.212>
6. Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*, 42, 226. <https://doi.org/10.1007/s10916-018-1082-9>
7. Pang, S., Yu, Z., & Orgun, M. A. (2017). A novel end-to-end classifier using domain transferred deep convolutional neural networks for biomedical images. *Computer Methods and Programs in Biomedicine*, 140, 283–293. <https://doi.org/10.1016/j.cmpb.2016.12.019>
8. Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35, 303–312. <https://doi.org/10.1016/j.media.2016.07.007>
9. Korbar, B., Olofson, A. M., Mirafior, A. P., Nicka, C. M., Suriawinata, M. A., Torresani, L., Suriawinata, A. A., & Hassanpour, S. (2017). Deep learning for classification of colorectal polyps on whole-slide images. *Journal of Pathology Informatics*, 8, 30. https://doi.org/10.4103/jpi.jpi_34_17
10. Yu, F., Sun, J., Li, A., Cheng, J., Wan, C., & Liu, J. (2017). Image quality classification for DR screening using deep learning. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 664–667). IEEE. <https://doi.org/10.1109/EMBC.2017.8036912>
11. Wong, K. C. L., Syeda-Mahmood, T., & Moradi, M. (2018). Building medical image classifiers with very limited data using segmentation networks. *Medical Image Analysis*, 49, 105–116. <https://doi.org/10.1016/j.media.2018.07.010>
12. Lakhani, P., Gray, D. L., Pett, C. R., Nagy, P., & Shih, G. (2018). Hello world deep learning in medical imaging. *Journal of Digital Imaging*, 31, 283–289. <https://doi.org/10.1007/s10278-018-0079-6>
13. Yasaka, K., Akai, H., Kunimatsu, A., Kiryu, S., & Abe, O. (2018). Deep learning with convolutional neural network in radiology. *Japanese Journal of Radiology*, 36(4), 257–272. <https://doi.org/10.1007/s11604-018-0726-3>

14. Bakator, M., & Radosav, D. (2018). Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3), 47. <https://doi.org/10.3390/mti2030047>
15. Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis—A survey. *Pattern Recognition*, 83, 134–149. <https://doi.org/10.1016/j.patcog.2018.05.014>
16. Chen, M. C., Ball, R. L., Yang, L., Moradzadeh, N., Chapman, B. E., Larson, D. B., Langlotz, C. P., Amrhein, T. J., & Lungren, M. P. (2018). Deep learning to classify radiology free-text reports. *Radiology*, 286(3), 845–852. <https://doi.org/10.1148/radiol.2017171115>
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3462–3471). IEEE. <https://doi.org/10.1109/CVPR.2017.369>
18. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*. <https://arxiv.org/abs/1711.05225>
19. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
20. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. <https://doi.org/10.1001/jama.2016.17216>