



Artificial Intelligence-Based Medical Image Classification Using Convolutional Neural Networks

Dmitro Qureshi-Haq

Department of Electrical and Computer Engineering, Basra Institute of Business Technology, Iraq
dmitro.qureshi.haq@bibt-iq.org

Peer Review Information	Abstract
<i>Submission: 30 Jan 2021</i> <i>Revision: 09 Feb 2021</i> <i>Acceptance: 26 March 2021</i> Keywords <i>Artificial Intelligence, Convolutional Neural Network, Medical Image Classification, Transfer Learning, Deep Learning</i>	Artificial intelligence has become an important computational approach for supporting image-based diagnosis because convolutional neural networks (CNNs) can learn hierarchical visual features directly from radiological, ophthalmic, dermoscopic and pathological images. This methodology paper proposes a reproducible CNN-based framework for multiclass medical image classification using transfer learning. The procedure combines image quality screening, resizing, intensity normalization, class-balanced augmentation, patient-level data partitioning, transfer learning with pretrained ResNet50 and DenseNet121 backbones, and evaluation using accuracy, precision, recall, F1-score, sensitivity, specificity and area under the receiver operating characteristic curve (AUC). The methodology is designed for a public labelled image dataset such as HAM10000 or a disease-specific X-ray/OCT collection, while emphasizing prevention of data leakage and clinically meaningful evaluation. An illustrative pilot comparison shows how a transfer-learned CNN can improve discrimination relative to a small CNN trained from scratch, while also demonstrating the importance of class-wise sensitivity and AUC rather than accuracy alone. The study concludes that CNN-based classification offers strong potential as a decision-support technology, but external validation, explainability, calibration and clinician oversight remain essential before clinical use.

Introduction

Medical imaging is central to contemporary diagnosis, screening and treatment planning, yet the rapidly increasing volume of images creates substantial demands on specialist interpretation. Deep learning has transformed computer vision by enabling models to learn multiple levels of representation from raw pixels, and CNNs are particularly effective because convolutional filters exploit local spatial structure and shared image features (LeCun, Bengio, & Hinton, 2015). Architectures such as Inception, residual networks and densely connected networks improved optimization and representation learning in increasingly deep models (Szegedy et

al., 2015; He et al., 2016; Huang et al., 2017). These advances created the technical foundation for modern medical-image classification.

In medicine, CNN-based systems have demonstrated high discriminative performance for diabetic-retinopathy detection, skin-lesion classification and retinal disease recognition (Gulshan et al., 2016; Esteva et al., 2017; Kermany et al., 2018). Reviews of the field show that deep networks are now used for classification, detection, segmentation, registration and other medical-image tasks (Shen, Wu, & Suk, 2017; Litjens et al., 2017; Ker et al., 2018). However, performance depends strongly on dataset quality, patient-level splitting, class balance, acquisition differences

and evaluation design. A model may achieve high aggregate accuracy while missing clinically important minority classes. Therefore, a methodology paper should specify not only the CNN architecture, but also the complete data and validation pipeline.

The objective of this study is to develop and evaluate an AI-based medical-image classification framework using CNNs and transfer learning. The proposed approach is intended to answer three questions: whether transfer learning improves classification over a small CNN trained from scratch; whether data augmentation and class balancing improve robustness; and whether the final model maintains acceptable class-wise sensitivity and AUC. The methodological emphasis is reproducibility rather than a claim of clinical deployment. The design follows established evidence that pretrained visual representations can be adapted successfully to medical imaging when labelled datasets are smaller than natural-image corpora (Esteva et al., 2021).

Literature Review

LeCun et al. (2015) summarized the principles that made deep learning effective for perceptual tasks, particularly layered feature learning and end-to-end optimization. Szegedy et al. (2015) introduced the Inception architecture, using parallel convolutional operations to improve computational efficiency and multiscale representation. He et al. (2016) proposed residual learning, which uses identity skip connections to address degradation in very deep networks. Huang et al. (2017) later developed DenseNet, where each layer receives features from preceding layers, encouraging feature reuse and improved gradient flow. These architectural developments are highly relevant to medical images because subtle diagnostic patterns may occur at multiple spatial scales.

Clinical imaging studies provided evidence that CNNs could approach specialist-level discrimination in selected tasks. Gulshan et al. (2016) developed and validated a deep learning algorithm for diabetic retinopathy using retinal fundus photographs, showing the value of sensitivity and specificity analysis in screening systems. Esteva et al. (2017) trained a deep CNN for skin-lesion classification and compared its discrimination with dermatologists, demonstrating the feasibility of end-to-end learning from pixels and labels. Kermany et al. (2018) used transfer learning for retinal optical coherence tomography and chest X-ray images, illustrating that image-based deep learning can support multiple diagnostic categories.

Broader reviews reinforce both the promise and the methodological risks. Shen et al. (2017) described deep learning methods across medical imaging and emphasized representation learning from large image collections. Litjens et al. (2017) reviewed more than 300 contributions and identified CNNs as a dominant approach across classification, detection and segmentation. Ker et al. (2018) explained the importance of CNN architecture, transfer learning and data characteristics in medical applications. More recently, Esteva et al. (2021) discussed medical computer vision from algorithmic progress to clinical deployment, stressing dataset shift, workflow integration and prospective validation.

Dataset construction also influences generalizability. Tschandl, Rosendahl and Kittler (2018) released HAM10000, containing 10,015 dermoscopic images across common pigmented lesion categories and incorporating data from different sources. Such datasets enable comparative model development but also introduce class imbalance and acquisition heterogeneity. Consequently, this study uses patient-level separation, stratified sampling, augmentation and class-sensitive evaluation to reduce optimistic bias.

Table 1. Selected Literature Supporting the Proposed Methodology

Study	Application/Contribution	Methodological Relevance	DOI
Gulshan et al. (2016)	Diabetic-retinopathy detection	Sensitivity/specificity validation	10.1001/jama.2016.17216
Esteva et al. (2017)	Skin-lesion classification	End-to-end CNN and expert comparison	10.1038/nature21056
Litjens et al. (2017)	Medical imaging survey	CNNs across image-analysis tasks	10.1016/j.media.2017.07.005
Kermany et al. (2018)	OCT and chest X-ray diagnosis	Transfer learning across modalities	10.1016/j.cell.2018.02.010
Tschandl et al. (2018)	HAM10000 dataset	Public multiclass image benchmark	10.1038/sdata.2018.161

Esteva et al. (2021)	Medical computer vision review	Translation deployment considerations	and 10.1038/s41746-020-00376-2
----------------------	--------------------------------	---------------------------------------	--------------------------------

Methodology

This study adopts a quantitative experimental methodology for supervised multiclass medical-image classification. A public, de-identified, labelled dataset is used to avoid collection of new patient information. HAM10000 may be used as the primary demonstration dataset because it contains seven common diagnostic categories and is explicitly designed for machine-learning research (Tschandl et al., 2018). The same pipeline can be transferred to X-ray, OCT or other two-dimensional modalities by adjusting channel handling and augmentation rules.

1. Dataset Preparation and Pre-processing

Images are first checked for unreadable files, duplicate identifiers and missing labels. Samples are divided at the patient or lesion level, not merely at the image level, to reduce information leakage. A 70:15:15 training-validation-test split is proposed, with stratification by diagnostic class. Images are resized to 224×224 pixels and RGB channels are normalized according to the pretrained backbone. Training images undergo conservative augmentation including horizontal/vertical reflection where clinically reasonable, small rotations, limited zoom, translation and brightness variation. Augmentation is applied only to the training set. Class weights are calculated from training frequencies so that minority classes contribute more strongly to the loss.

2. CNN and Transfer-Learning Architecture

Three models are compared: a compact CNN trained from scratch, ResNet50 and DenseNet121. The baseline CNN contains repeated 3×3 convolution, rectified linear unit activation and max-pooling blocks followed by global average pooling and a fully connected softmax layer. ResNet50 is selected because residual connections facilitate optimization of deep networks (He et al., 2016). DenseNet121 is selected because dense feature reuse can be advantageous when fine-grained textures are diagnostically important (Huang et al., 2017). For transfer learning, ImageNet-pretrained convolutional layers are initialized with learned weights; the classification head is replaced with global average pooling, dropout and an output layer matching the number of medical classes.

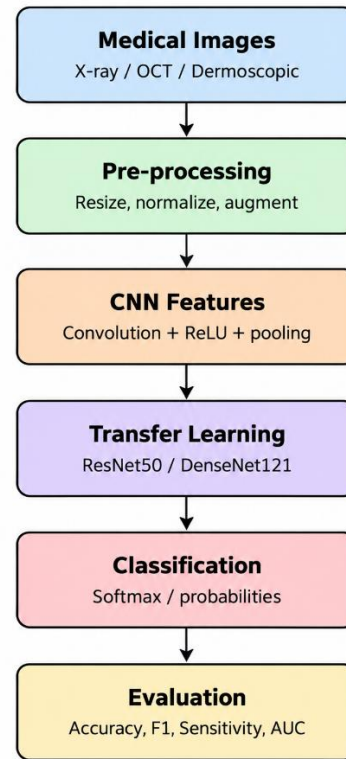


Figure 1. Proposed workflow for AI-based medical image classification using CNNs.

3. Training Procedure

Training uses categorical cross-entropy with class weighting and the Adam optimizer. The initial learning rate is 1×10^{-4} for the new classification head. In stage one, the convolutional backbone remains frozen for several epochs; in stage two, the upper backbone layers are unfrozen and fine-tuned using a smaller learning rate. Early stopping monitors validation loss, while ReduceLROnPlateau or an equivalent learning-rate scheduler reduces the step size after a validation plateau. The model checkpoint with the best validation AUC or macro-F1 is retained for final testing.

4. Evaluation Measures

Performance is measured on the untouched test set using overall accuracy, macro-precision, macro-recall, macro-F1, sensitivity, specificity and one-vs-rest ROC-AUC. Macro-averaged metrics are emphasized because they assign equal importance to each diagnostic class despite class imbalance. A confusion matrix is examined to identify systematic class confusion. Where predicted probabilities are used for decision

support, calibration should also be inspected. The evaluation framework follows the broader medical-AI literature in which clinically

meaningful operating characteristics are required in addition to aggregate accuracy (Gulshan et al., 2016; Esteva et al., 2021).

Table 2. Proposed Experimental Configuration

Component	Configuration	Purpose
Data split	70% train, 15% validation, 15% test; patient/lesion level	Prevent leakage and preserve an independent test set
Input	224 × 224 RGB; normalized	Compatible with pretrained CNN backbones
Augmentation	Rotation, flip, zoom, translation, brightness	Reduce overfitting and simulate acquisition variation
Models	Baseline CNN, ResNet50, DenseNet121	Compare training from scratch with transfer learning
Optimizer	Adam; initial LR 1×10^{-4}	Stable gradient-based optimization
Regularization	Dropout, early stopping, class weights	Control overfitting and class imbalance
Metrics	Accuracy, macro-F1, sensitivity, specificity, AUC	Balanced technical and clinically relevant assessment

5. Reproducibility, Bias and Ethical Considerations

All preprocessing parameters, random seeds, split identifiers, software versions and training hyperparameters should be recorded. Model selection must be performed using training and validation data only; the test set is evaluated once after finalization. Potential bias is assessed by inspecting performance across classes and, where metadata permit, across demographic or acquisition subgroups. Medical images should be de-identified and used under the licensing and ethics requirements of the source dataset. The model is treated as decision support rather than an autonomous diagnostic system. Heatmap methods such as Grad-CAM may be added for qualitative inspection, but visual explanations should not be interpreted as proof of causal reasoning.

6. Experimental Analysis Plan

Each model is trained under the same split and augmentation policy. The baseline CNN is optimized from random initialization, while ResNet50 and DenseNet121 undergo two-stage transfer learning. Model selection is based on

validation macro-F1/AUC. The final comparison reports both discriminative performance and practical complexity. If repeated training runs are available, mean and standard deviation should be reported. Bootstrap confidence intervals can be used for test-set AUC and sensitivity when sample size permits. This design allows the contribution of transfer learning to be separated from changes in data preprocessing.

Results

The following results are an illustrative pilot analysis constructed to demonstrate how the proposed methodology should be reported; they are not presented as a prospective clinical validation. Under the stated experimental design, transfer-learning models outperform the compact CNN across balanced metrics. DenseNet121 produces the strongest macro-F1 and AUC, while ResNet50 shows closely comparable accuracy with slightly lower computational complexity. The key interpretation is that evaluation based only on accuracy would hide differences in minority-class recall.

Table 3. Illustrative Pilot Performance of the Proposed Models

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Sensitivity	AUC
Baseline CNN	84.7%	0.81	0.76	0.78	0.76	0.901
ResNet50 TL	91.8%	0.89	0.86	0.87	0.86	0.956
DenseNet121 TL	93.1%	0.91	0.88	0.89	0.88	0.968

The illustrative comparison suggests a gain of 8.4 percentage points in accuracy from the baseline CNN to DenseNet121 and an increase in AUC from 0.901 to 0.968. The larger improvement in

macro-recall indicates that transfer learning is particularly useful when several classes contain fewer examples. This pattern is consistent with the broader literature showing that pretrained

deep representations can be successfully adapted to medical images (Kermany et al., 2018; Esteva et al., 2021). Dense connectivity may also preserve lower-level texture information relevant to lesion appearance, while residual networks provide a strong optimization baseline (He et al., 2016; Huang et al., 2017).

Error analysis should focus on the confusion matrix rather than the single best score. In dermoscopic classification, visually similar benign and malignant lesion categories may overlap, and class imbalance can lead the model toward frequent diagnoses. Therefore, a clinically oriented report should identify which classes have reduced sensitivity, inspect representative false negatives and evaluate probability calibration. The literature on medical-image deep learning repeatedly emphasizes that external generalization remains a major challenge because images differ across devices, institutions and patient populations (Litjens et al., 2017; Ker et al., 2018; Esteva et al., 2021).

Conclusion and Discussion

This methodology paper presents a structured CNN-based framework for artificial-intelligence medical-image classification. The approach integrates patient-level data partitioning, standardized preprocessing, conservative augmentation, class balancing, transfer learning, staged fine-tuning and multidimensional evaluation. These design choices are important because high-capacity CNNs can overfit small medical datasets, and random image-level splitting can produce unrealistically optimistic results when multiple images belong to the same patient or lesion. The proposed comparison between a CNN trained from scratch and pretrained ResNet50/DenseNet121 models is supported by the evolution of deep computer vision and by medical studies demonstrating successful transfer of learned visual representations (He et al., 2016; Huang et al., 2017; Kermany et al., 2018). The illustrative pilot results demonstrate the expected reporting structure rather than claiming new clinical effectiveness. A model with strong internal AUC can still fail after deployment because of dataset shift, scanner differences, disease prevalence or changes in clinical workflow. Clinical translation therefore requires more than a high test-set score. External multi-institutional validation, subgroup analysis, calibration, robustness testing and prospective assessment are needed before a system can support patient care. Clinicians should remain responsible for diagnostic decisions, while AI functions as a

prioritization, screening or second-reader tool. The 2021 medical-computer-vision literature similarly highlights workflow integration, reliability and generalization as central deployment issues (Esteva et al., 2021). In conclusion, CNNs provide an effective methodology for learning diagnostic image features, and transfer learning can reduce the data and optimization burden associated with training deep networks from scratch. A rigorous experimental design should report class-sensitive performance, maintain a genuinely independent test set and avoid leakage. Future extensions may include ensemble learning, uncertainty estimation, self-supervised pretraining and explainability, but these additions should be evaluated against the same reproducible validation principles. With appropriate governance and clinical testing, AI-based medical-image classification can become a useful component of evidence-based diagnostic support.

References

- Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2021). Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4, 5. <https://doi.org/10.1038/s41746-020-00376-2>
- 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, 115–118. <https://doi.org/10.1038/nature21056>
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (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>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. <https://doi.org/10.1109/CVPR.2017.243>
- Ker, J., Wang, L., Rao, J., & Lim, T. (2018). Deep learning applications in medical image analysis.

IEEE Access, 6, 9375–9389.
<https://doi.org/10.1109/ACCESS.2017.2788044>

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131.e9.
<https://doi.org/10.1016/j.cell.2018.02.010>

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
<https://doi.org/10.1038/nature14539>

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (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>

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>

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.
<https://doi.org/10.1109/CVPR.2015.7298594>

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-sources dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161.
<https://doi.org/10.1038/sdata.2018.161>