ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

TOWARD EXPLAINABLE AND RESOURCE-EFFICIENT MEDICAL IMAGE CLASSIFICATION A CRITICAL REVIEW AND PROPOSED HYBRID CNN–TRANSFORMER WEB FRAMEWORK

Muhammad Mubeen Goraya
BSCS (Scholar), Virtual University of Pakistan
bc240221233mmg@vu.edu.pk
Corresponding author
Muhammad Kaleemullah
Lecturer
Virtual University of Pakistan
Email: m.kaleem@vu.edu.pk
Muhammad Usman
University of Engineering and Technology Lahore
2022cs433@student.uet.edu.pk
Ali Awais
aawaisdr@gmail.com
MBA IT management AIOU
MBA banking and finance Coventry University, UK

ABSTRACT
Despite the significant progress made by deep learning in image classification, there are four common problems that limit the potential of practical deployment of medical image classifiers: convolutional neural networks (CNNs) excel at modelling locally available texture but provide weak global context; vision transformers offer strong global context but require a lot of data and computation; many state-of-the-art classifiers are barely interpretable; and research prototypes have not been transferred to secure, responsive web applications. This paper reviews and summarizes the relevant literature that was identified in a BSCS image-classification project, and it adds basic literature on residual networks, lightweight CNNs, hierarchical transformers, multiscale attention, and gradient-based explanations, which are all technically relevant. Based on this synthesis, it introduces a resource-efficient hybrid system designed to construct multi-resolution feature maps with a lightweight pre-trained CNN, to model cross-region dependencies in a compact multiscale transformer, to combine the features for classification in a gated manner, and to synthesize complementary explanations using Grad-CAM and attention-rollout. There is a web deployment architecture defined, with image validation, privacy-sensitive pre-processing, model inference, calibrated confidence reporting, explanation overlays, and auditing. A stratified cross-validation protocol, imbalance-aware metrics, calibration, latency, model size, ablation studies and statistical comparisons with CNN and transformer baselines are also included in the paper, as they are reproducible. It is not just a precision claim that will come to the fore, but a coherent and actionable vision that connects model design, explainability, resource efficiency, and web deployment. The framework aims to facilitate future empirical application to a well-documented medical image data set, and the achievement of a closer link between classification research and deployable software for aiding in decision making.

Keywords: *medical image classification; convolutional neural network; vision transformer; multiscale feature fusion; transfer learning; explainable artificial intelligence; Grad-CAM; web application*

1. INTRODUCTION

1. Introduction

A fundamental computer vision problem is image classification, in which a computational model is used to classify an input image into one of the given categories. In medical imaging, classification can be used to assist in screening, triage, education, quality assurance, and

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

computer-assisted diagnosis in modalities like dermoscopy, radiography, computed tomography, magnetic resonance imaging, and microscopy. These applications have spurred rapid advances in deep learning, especially CNNs, which learn hierarchical visual features directly from the pixels without requiring any manually engineered feature descriptors.

CNN architectures like ResNet and DenseNet facilitated optimization and feature reuse with residual and dense connectivity, respectively, while MobileNetV2 and EfficientNet opened up the possibility of doing image recognition with limited computational resources (He et al., 2016; Huang et al., 2017; Sandler et al., 2018; Tan & Le, 2019). Transfer learning also minimized the amount of data needed for training large models, by fine-tuning representations learned from big natural-image corpora for smaller domain-specific ones. The project literature also mentions transfer learning as a frequent approach for medical, satellite, scene and tree disease classification.

But even then the representation problem is not completely addressed by convolution. CNNs are helpful for learning local patterns, but relationships between distant image regions might take several layers or special care. To overcome this, the Vision Transformer (ViT) models, which propose to use image patches as tokens and compute self-attention over the entire image to incorporate global dependencies (Dosovitskiy et al., 2021). Hierarchically and multiscale designs like Swin Transformer and Multiscale Vision Transformers are devised to boost efficiency and achieve valuable feature pyramids for fine-grained visual recognition (Fan et al., 2021; Liu et al., 2021). Therefore, the study of multiscale transformers and CNN-transfomer hybrids have been increasingly explored in the recent medical-image research.

Model performance is not enough for responsible deployment. Medical data frequently suffer from class boundaries, an imbalance of datasets, artefacts from acquisition, and institution-specific biases. A model can have a high test accuracy and be poorly calibrated or sensitive to distribution shift or targeted to irrelevant regions. Explainability methods, like Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017), can be used to find out what spatial regions have been contributing to a prediction; however, explanation maps cannot be interpreted as evidence of causal reasoning (Selvaraju et al., 2017).

Another gap is between model development and software delivery. A research notebook can perform a successful classification of images, but a usable web application also needs validation of the input, deterministic processing of the input, secure handling of files, versioning of the model, controlling for the latency, accessible output, error handling, and privacy-conscious logging. The end-of-course project that ignited this article is specifically oriented towards an image classification web application, and the literature reviewed focused on deep learning, transfer learning, multiscale attention and python based web systems. These components can be put into a unified deployable structure.

This paper instead tries to rephrase the project material as a critical review and proposed technical architecture and not as an empirical performance paper. Its contributions are fourfold: firstly, it filters and synthesizes the relevant technically relevant studies from the project corpus; secondly, it identifies the unresolved trade-off between detail at the local level, context at the global level, efficiency and interpretability; thirdly, it proposes a compact hybrid CNN-transformer architecture and web inference workflow; and fourthly, it defines an evaluation protocol that would allow the design to be converted into a reproducible experimental study once a documented dataset and implementation are provided.

2. REVIEW DESIGN

2.1 Source Corpus and Relevance Screening

The source project abstracted 20 papers (or technical records) which were primarily published in the years 2018-2025. The range of topics covered by the corpus was broad, with some entries focused on medical image classification, while others were related to other areas of AI like predictive modeling, signal integrity, or quantum computing. The authors chose only to include records that had a direct contribution to image classification, transfer learning, medical imaging,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

explainability or web-based diagnosis to avoid artificially expanding the evidence base. There were eight project records that satisfied that criterion of relevance. The synthesis was then reinforced by the basic primary papers on the particular architectures employed in the proposed structure.

This is not a formal systematic review or meta-analysis, but rather an organized critical review. Effect sizes, independent screening, risk-of-bias assessment, and database-specific search strings, duplicate removal, were not documented in the source project. Hence, the present paper does not apply the PRISMA terminology and does not combine performance numbers from data sets with different classes, number of samples, preprocessing and validation design.

2.2 Analytical Themes

The studies were organized into five analytical themes: automatic feature learning using CNNs, transfer learning using limited labels, multi-scale and transformer-based context modeling, explainability and practical web deployment. The analysis was not directed towards direct comparisons of reported accuracies due to the fact that the accuracies reported may not be comparable as different studies use different datasets and split protocols.

3. CRITICAL SYNTHESIS OF EXISTING APPROACHES

Study / architecture	Primary contribution	Strength relevant to this project	Limitation or caution
He et al. (2016), ResNet	Residual learning for deep CNN optimization	Strong transferable backbone and stable fine-tuning	Higher variants can be expensive for web inference
Huang et al. (2017), DenseNet	Dense feature reuse and short gradient paths	Efficient reuse of low- and high-level features	Memory consumption grows with dense connections
Sandler et al. (2018), MobileNetV2	Inverted residuals and linear bottlenecks	Low parameter count and deployment suitability	May lose fine-grained accuracy on difficult classes
Tan and Le (2019), EfficientNet	Compound scaling of depth, width, and resolution	Balanced accuracy-efficiency trade-off	Scaling choices remain dataset dependent
Dosovitskiy et al. (2021), ViT	Global self-attention over image patches	Long-range dependency modeling	Data hungry and computationally demanding
Fan et al. (2021), MViT	Hierarchical multiscale transformer	Progressive token pooling and multiresolution context	More complex implementation and tuning
Liu et al. (2021), Swin	Shifted-window hierarchical attention	Local efficiency with cross-window interaction	Window configuration influences results
Selvaraju et al. (2017), Grad-CAM	Gradient-based localization explanations	Model-agnostic spatial explanation for CNN-based classifiers	Coarse maps can be unstable or overinterpreted

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Study / architecture	Primary contribution	Strength relevant to this project	Limitation or caution
Hu et al. (2025)	Multiscale transformer for skin-lesion classification	Demonstrates the value of local-global fusion and visual explanations	Evidence is dataset specific and requires external replication
Ahmad et al. (2025); Chen (2025)	Reviews of CT and CNN medical classification	Identify transfer learning, data quality, and explainability as recurring themes	Review evidence inherits heterogeneity of underlying studies

3.1 CNNs, Transfer Learning, and Feature Hierarchies

CNNs are given as useful because of the encoding of locality and translation-related prior knowledge that is useful when the medical dataset is smaller than the natural-image corpora. Early layers are learning edges and textures, while deeper layers are learning more and more class specific structures. Deeper networks can optimize using residual connections without having to learn a full transformation from each stacked block, and dense connectivity enables feature re-use at different levels.

The key practical benefit of transfer learning is not merely the similarity of the natural and medical images, but that early and intermediate visual forms are provided as a useful starting point. Fine tuning must also be done with care, however. Freezing the whole backbone could be suboptimal for domain-specific structures, and high learning rates in all layers could be detrimental to the transfer of representations. To this end, a sequential schedule – training the classifier head, unfreezing the upper blocks of the backbone network, followed by applying low-rate end-to-end fine-tuning – is preferred for the proposed system.

A web application is particularly interested in lightweight architectures. The MobileNetV2 network employs depthwise separable convolutions and inverted residual bottlenecks to minimize computation, and the EfficientNet network has been designed to trade off resolution of the input, depth of the network, and width of the network. The use of transformer layers in a hybrid approach is ideal for using them only on a small, multi-scale feature map, and the backbones provide a convenient feature extractor for this hybrid model.

3.2 Transformers and Multiscale Context

Self-attention allows each token to sum over other tokens based on learned relevance. This property may be used to link regions that are far away but are yet jointly informative in image classification. In contrast, Standard ViT suffers from quadratic attention complexity, and is usually trained at a large scale. Hierarchical designs relieve this burden, both in that they restrict the spatial attention and that they pool tokens and/or operate on increasingly coarse feature maps.

Multiscale modelling is crucial if a class requires a global configuration and a local texture. There may be different spatial resolution for representing a lesion boundary, a pulmonary opacity or a cellular morphology. One patch size may be able to reduce the size of small details or break down larger structures. This is a technically valid direction for the project and the researchers are going the right way to achieve it, but the implementation is resource-wise valid instead of using a full transformer, which compromises the responsiveness of the web.

3.3 Explainability, Calibration, and Responsible Interpretation

Grad-CAM derives a class-specific heatmap based on gradients that are passing through a chosen convolutional feature layer. It is appealing in web interfaces since it generates a user friendly overlay without having to retrain models. In Hybrid, Grad-CAM could be applied to CNN branch, and attention rollout could summarize the interactions between tokens in transformer. Consensus

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

between two views can help to boost interpretive confidence, while significant differences could lead to an alert for manual interpretation.

Proper probability values should not be presented as a kind of gospel. Distributed shift and class imbalance can cause deep neural networks to become overconfident. Calibration error, and temperature scaling if needed, should then be part of the evaluation protocol on a held out validation set. The output of the interface should be referred to as model confidence, not clinical certainty.

3.4 From a Trained Model to a Web Application

The project literature contains a simple web system for disease diagnosis using Python, for the sake of domain users. A modern image-classification service should be a presentation layer from an inference service. A separate backend API is responsible for the file validation, preprocessing, inference, explanation generation and response serialization, while the selection of images and the visualization of the results are handled by the browser, and the model artifact itself is versioned separately. This separation allows for better maintainability, and allows for a different classifier to be replaced without having to change the user interface.

4. RESEARCH GAP AND DESIGN REQUIREMENTS

There is no literature that shows a single architecture that satisfies all of the above requirements, namely fine-detail sensitivity, global context, low inference cost, calibrated confidence, interpretable output, and secure web deployment. One of the downside of using high capacity CNNs is that they tend to be expensive. With lightweight CNNs, the fine-grained distinctions may be lost. Transformers add context to global data but add data and computational demands. Usually, the explanation is added in after the model has been developed, or the web deployment details are not included.

The proposed framework is therefore guided by the following requirements:

- Local-detail preservation through a pretrained convolutional feature pyramid.
- Global and cross-scale reasoning through a compact transformer operating on reduced tokens.
- Resource efficiency appropriate for server-side web inference and possible edge deployment.
- Class-imbalance awareness and calibrated probability reporting.
- Dual explainability through Grad-CAM and transformer attention rollout.
- Privacy-aware upload handling, model versioning, reproducible preprocessing, and audit metadata.
- Modularity so that datasets, backbones, and classification heads can be replaced without redesigning the full application.

5. PROPOSED HYBRID CNN–TRANSFORMER FRAMEWORK

5.1 System Overview

The proposed pipeline is shown in figure 1. Uploaded image is going through validation and deterministic preprocessing. The feature maps are generated by a light neural network at multiple resolutions.The light CNN generates feature maps at multiple resolutions. The features maps are mapped to a common embedding dimension and mapped to a short token sequence. Transformer encoders represent relationships within and across scales. A gated fusion layer is applied before the final classifier, which fuses the convolutional representation and a transformer representation. The prediction service returns explanation overlays, confidence and uncertainty indicators as well as a class.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

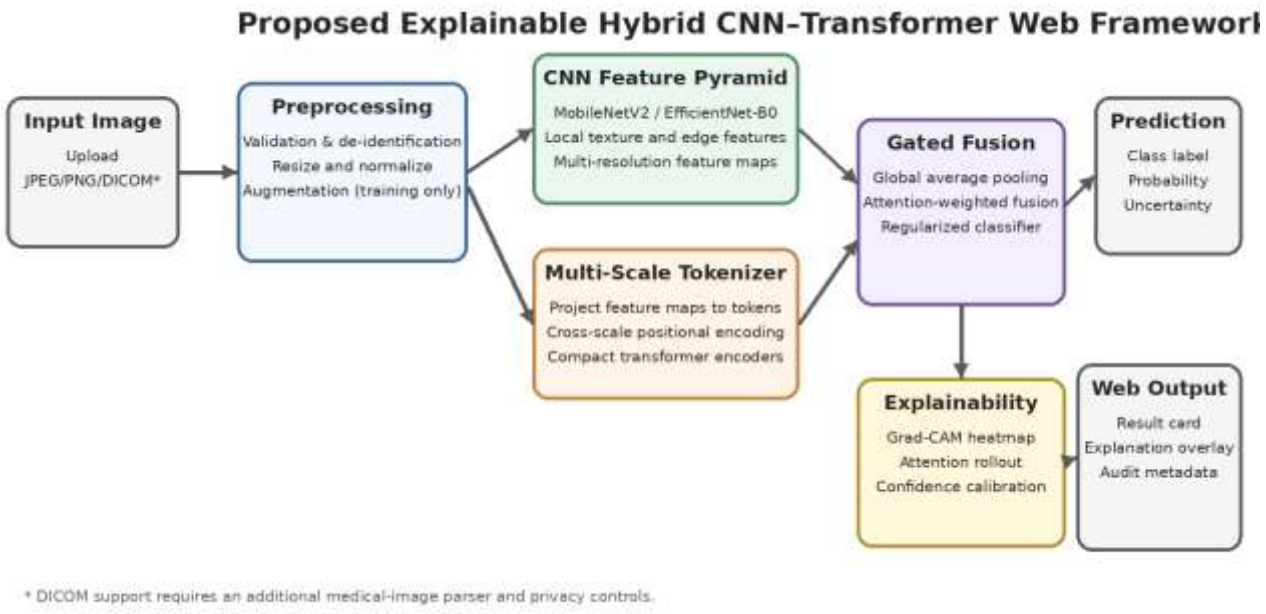


Figure 1. Proposed resource-aware hybrid CNN–Transformer architecture and web inference workflow.

5.2 Input Validation and Preprocessing

All the preprocessing steps should be the same during training and inference, except for stochastic augmentation. The web service needs to check the file type, file signature, image size and maximum size first. Any file that has not been supported or is corrupted should be rejected prior to decoding. Medical metadata should not be stored if it is not needed and DICOM images should be de-identified before using.

Images can be scaled to 224 × 224 size, be converted to the channel format expected by the pretrained backbone, and be rescaled to the statistics of the backbone using the statistics expected by the pretrained backbone. Some examples of training augmentation include bounded rotation, translation, scale variation, flipping in the horizontal direction when anatomically appropriate, slight color and/or contrast variation, and random erasing. Augmentations have to be approved by domain experts because natural image transformations that do not affect the image may affect the meaningful image orientation or image intensity for medical images.

5.3 Convolutional Feature Pyramid

For the first implementation, the suggested backbone is MobileNetV2 or EfficientNet-B0. The intermediate output of three stages is synthesized in a feature pyramid for fine, intermediate and semantic information. Suppose the three feature maps are denoted by F1, F2, and F3. A 1 × 1 projection projects their channel dimensions on to a common embedding size d. Adaptive pooling limits the spatial sizes and hence the number of tokens fed into the transformer.

5.4 Multiscale Tokenization and Attention

A feature map is taken as a flattened vector of tokens and fed with a scale embedding and positional encoding for each feature map. The token set is denoted as Z = [Z1; Z2; Z3] and each Zn represents a CNN scale. A compact transformer encoder is used to implement layer normalization, multi-head self-attention, residual connections, and feed-forward network. Cross-scale attention enables a token, representing a small local pattern, to relate to tokens, representing larger anatomical context.

Scaled dot-product attention is $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})V$, where Q, K, and V are the projections obtained from the input tokens and d_k is the dimension of the keys. To make inference practical, the number of encoder blocks is first kept to 2 to 4, and the number of attention heads and embedding size are tuned by validation and not taken from a large ViT.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

5.5 Feature Fusion and Classification with Gates.

The CNN branch is summarized by a global average pooling, and the transformer branch is summarized by attention pooling or a learned classification token. A gating network is used to learn the contribution of each branch relative to the others. A fused representation h is then obtained after dimensional alignment by computing a gate $g = \text{sigmoid}(W[c;t] + b)$, followed by dimensional alignment, and then computing $h = g \odot c + (1 - g) \odot t$. The final linear classifier is made less prone to overfitting by dropout and weight decay.

5.6 Training Strategy

1. Initialize the CNN backbone with publicly available pretrained weights and randomly initialize transformer, fusion, and classification layers.
2. Train the new layers while the backbone is frozen, using class-weighted cross-entropy or focal loss if imbalance is substantial.
3. Unfreeze the upper CNN stages and fine-tune them with a lower learning rate than the newly added layers.
4. Optionally perform low-rate end-to-end fine-tuning with early stopping based on validation macro-F1 or AUC rather than training accuracy.
5. Select the final checkpoint using a prespecified validation criterion and calibrate probabilities on data not used to fit model weights.

AdamW is suitable for the transformer and fusion layers, while stochastic gradient descent with momentum can be evaluated for CNN fine-tuning. The optimizer should be treated as a tunable component and reported transparently with learning rates, schedules, batch size, weight decay, seed values, and stopping criteria.

5.7 Explainability and Uncertainty Output

The inference service generates a Grad-CAM heatmap from the last spatial CNN block and an attention-rollout map from transformer layers. Both maps are resized to the input resolution and returned as optional overlays. Explanations should be accompanied by a notice that they indicate model sensitivity rather than validated pathology localization. Confidence calibration, entropy, and a minimum-confidence threshold can support a defer-to-human rule for uncertain inputs.

Module	Recommended initial configuration	Purpose
Input	224 × 224 RGB or modality-specific channels	Standardize tensor dimensions
CNN backbone	MobileNetV2 or EfficientNet-B0, pretrained	Efficient local and hierarchical feature extraction
Feature taps	Three backbone stages	Create fine-to-coarse feature pyramid
Projection	1 × 1 convolution to shared d	Align channels for tokenization
Transformer	2–4 blocks; 4–8 heads; compact embedding	Model cross-region and cross-scale dependencies
Fusion	Gated CNN/transformer fusion	Adaptively combine inductive bias and global context
Classifier	Dropout + linear softmax head	Predict image category
Explainability	Grad-CAM + attention rollout	Provide complementary visual evidence

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Module	Recommended initial configuration	Purpose
Serving	FastAPI/Flask API with model versioning	Deliver controlled web inference

6. WEB APPLICATION ARCHITECTURE

6.1 Client Layer

The client interface shall offer drag and drop upload, indication of supported formats, preview, consent / privacy notice, progress feedback and result card. The result card should show the predicted class, calibrated confidence, model version, processing time, and optionally, an explanation overlay. The interface should not include any phrases like “diagnosis confirmed.”

6.2 Inference API

A Flask or FastAPI backend provides a versioned endpoint with a name like classify/1.The Flask or FastAPI backend provides a versioned endpoint like /api/v1/classify. The endpoint validates the request, creates an in-memory image object, applies a serialised pre-processing pipeline, inference is done in evaluation mode, explanations are generated and returned as a structured JSON. The default is to NOT store the contents of a file. When a log operation is performed, the system should include non-identifying information like the model version, latency, error code, and input dimensions.

6.3 Packaging and Reproducibility:

The trained model, class mapping, pre-processing parameters and configuration, calibration parameters, and software requirements should be provided in one versioned release. Export to ONNX might be useful for portability, but the numbers exported need to be verified with the original framework. Dependencies can be isolated through containerization for repeatability. A health-check endpoint should check that the service is loaded, and the model, but not expose private data.

6.4 Security and Privacy Controls

- Enforce file-size limits, MIME and signature checks, and safe image decoding.
- Reject executable, polyglot, malformed, or unsupported files.
- Strip metadata and avoid retaining uploads unless explicit consent and governance exist.
- Use HTTPS, request-rate limits, authentication where appropriate, and server-side timeouts.
- Record model versions and configuration hashes for traceability.
- Clearly state that the software is a research decision-support prototype, not a substitute for professional diagnosis.

7. REPRODUCIBLE EVALUATION PROTOCOL

7.1 Dataset Documentation and Splitting

The author should identify a legally usable dataset prior to empirical publication and report the source and license of the dataset, as well as its modality, class definitions, sample numbers, patient or subject structure, acquisition settings, exclusion conditions etc. For each patient or subject with more than one image, splitting shall be done at the patient/subject level. Near duplicate information can be leaked and optimistic results can be achieved when using image level random splitting.

If available, it is recommended to use stratified five-fold cross validation with one held-out external/temporal test set. Hyperparameters need to be determined in the training folds. The test set should not be modified until the architecture and decision thresholds are determined.

7.2 Baselines and Ablation Studies

The proposed model should be compared with a conventional CNN (e.g., ResNet50), a lightweight CNN (e.g., MobileNetV2, EfficientNet-B0), and a compact transformer/ViT baseline

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

trained using the same preprocessing and split conditions. Alongside predictive metrics, there should be a parameter count, floating-point operations, model size, and latency.

One by one, ablation studies should remove: multiscale inputs, transformer blocks, gated fusion, class weighting, calibration, and individual explanation mechanisms. These experiments seek to establish whether the proposed complexity really offers a measurable improvement without simply adding up the number of parameters.

7.3 Metrics

Metric	Why it is needed	Reporting recommendation
Accuracy	Overall proportion of correct predictions	Report only with class distribution
Macro precision / recall / F1	Weights each class equally	Primary metrics for imbalanced multiclass data
Sensitivity and specificity	Clinically interpretable error trade-off	Report per class or one-vs-rest
ROC-AUC	Ranking performance across thresholds	Use macro and classwise values
PR-AUC	More informative for rare positive classes	Report for imbalanced binary/one-vs-rest tasks
Confusion matrix	Shows systematic class confusions	Provide counts and normalized values
Expected calibration error	Assesses probability reliability	Report before and after calibration
Latency and throughput	Measures deployment feasibility	Report hardware, batch size 1, and warm-up policy
Parameters / model size	Measures resource burden	Compare all baselines consistently

7.4 Statistical Analysis

7.4 Statistical Analysis

Patient-level bootstrap resampling should be used to generate confidence intervals. For macro-F1, paired bootstrap tests could be used to compare it, and for one-vs-rest or binary evaluation, DeLong's method could be used to compare ROC-AUC. A modified significance procedure should be used for multiple model comparisons. The manuscript should describe the variations between folds and random seeds and not a single good run of the experiment.

7.5 Robustness and Generalization Tests

- Evaluate performance across acquisition devices, sites, demographic subgroups, and image-quality strata when metadata permit.
- Test bounded perturbations such as compression, blur, brightness shift, and moderate cropping.
- Assess out-of-distribution behavior using images from unrelated classes or a different dataset.
- Inspect misclassified cases and explanation maps with domain expertise rather than reporting aggregate metrics alone.
- Measure calibration drift and define a retraining or monitoring policy for deployment.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

8. DISCUSSION

The proposed architecture is based on the idea that transformers do not always perform better than CNNs. The CNN branch adds a strong locality bias with a computationally efficient feature hierarchy. The transformer branch is used after spatial compression, enabling to model long-range and cross-scale dependencies without processing the full-resolution image as a large token sequence. The gated fusion component gives a mechanism to the model to be more dependent on local evidence or more dependent on contextual evidence on different inputs.

This design echoes the project literature where transfer learning, multiscale feature extraction, attention and explainability have been mentioned repeatedly as directions of importance. In addition, it overcomes a limitation of the source project: The accuracy presented in many of the summarized papers is high but they fail to supply a commonly used dataset or a validation procedure that would allow direct comparisons. Patient-level splitting, class-balanced metrics, calibration, ablation and latency are thus the first priorities of the present evaluation plan.

The web layer isn't merely a decorative layer. Whether a classifier works as it did during experimentation or not is a direct consequence of the following factors: reproprocessable preprocessing, input security, model versioning, and privacy controls. The separation of the API, model artifact and user interface allows for controlled updates and external auditing. For medically oriented images, the application must be clearly identified as a research version of the software, until it has been clinically validated, shown to be FDA/Center for Devices and Radiological Health (CDRH) compliant (where applicable), and has been subjected to prospective evaluation.

A successful empirical demonstration would validate three points – that multiscale transformer fusion produces better predictive performance than lightweight CNN baselines; that the benefit outweighs the increased computational burden; and that explanation maps are always focused on the task-relevant regions. Answers to these questions should be arrived at experimentally, not assumed from the architecture.

9. LIMITATIONS

There are some weaknesses in this manuscript. First, it is based on a final year project document that includes literature summaries and a proposed approach, but no inspectable code, dataset, experiment log or result table. So it cannot be said to be accurate, valid, or perform in the real world. Second, the review is not systematic because there was poor documentation of the original search and screening process to enable replication. Third, the proposed model is still dataset agnostic, where optimal architecture, augmentation or loss function will be dependent on the modality and class definitions. Fourth, explanation techniques like Grad-CAM, attention rollout do not ensure causal reasoning and are susceptible to layer selection and pre-processing. Lastly, medical applications need to be checked for domain, validated externally, governed by privacy and have a clear human oversight process.

10. Conclusion

In this article, a final-year project involving image classification is converted into a logical and understandable review and design document. Transfer learning is supported by literature as an efficient starting point, multiscale attention as a mechanism to combine fine and global evidence and explainability as an important component of responsible image classification. The proposed system consists of a light-weight CNN feature pyramid, small multi-scale transformer encoders, gated fusion, calibrated predictions, complementary explanation maps, and a modular web service. Now it is up to the clear empirical implementation to establish its scientific value. In future, a licensed dataset should be selected, subject-level separation should be maintained, strong baselines should be compared, ablation and calibration should be performed and latency and uncertainty should be reported in addition to classification metrics. If followed that protocol would yield the evidence to follow the format of a regular experimental research article.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

TOWARD EXPLAINABLE AND RESOURCE-EFFICIENT MEDICAL IMAGE CLASSIFICATION

A Critical Review and Proposed Hybrid CNN–Transformer Web Framework

Muhammad Mubeen Goraya (BSCS Scholar, Virtual University of Pakistan, bc240221233mmg@vu.edu.pk)
Muhammad Kaleemullah (Lecturer, Virtual University of Pakistan, m.kaleem@vu.edu.pk) - **Corresponding Author**
Muhammad Usman (University of Engineering and Technology Lahore, 2022cs433@student.uet.edu.pk)
Ali Awais (MBA IT Management AIOU & Banking and Finance Coventry University, UK, aawaisdr@gmail.com)

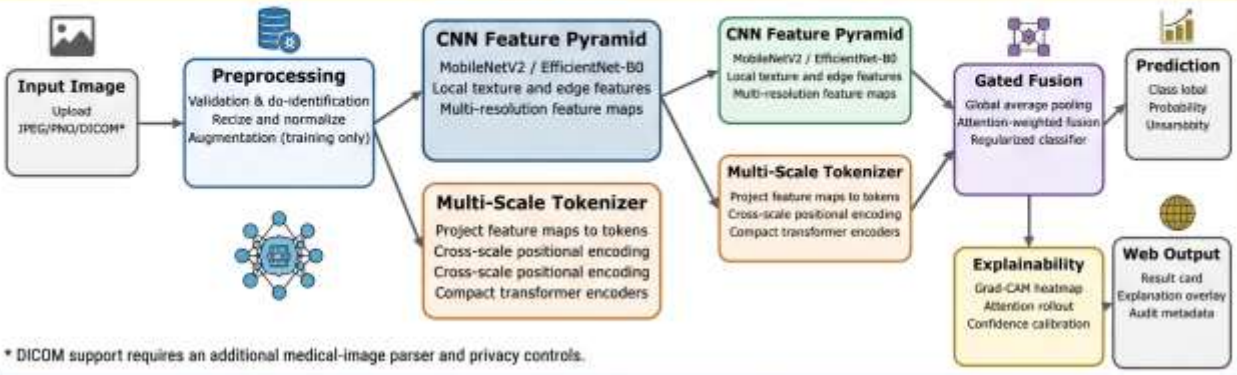
ABSTRACT

Resource-efficient, interpretable, medical and interial image classifiers is usuarlly recoed to preparing lightweight CNNs and united lightweight CNNs and multiscale transioate-masgrosment, united multiscale transformers into a secure compatible, onm-est mobcale lightweight CNNs and multiriscale nrachament and multiscale transformers inbodly-nending a secure wnd secure web framework.

RESEARCH GAP & DESIGN REQUIREMENTS

- **Local-detail preservation**, innnovating innnest and prevɪn medical image classifiers
- **Global reasoning**, globa nɛsernosation, **low inference cost**, and cɪlbrated rathoaility
- **Calibrated confidence**, aligrated calibrated confidence, and consetcant is soon sesability

PROPOSED HYBRID FRAMEWORK



* DICOM support requires an additional medical-image parser and privacy controls.

WEB APPLICATION ARCHITECTURE & SECURITY

- Client-server separation, esource-efficient
- Flask/FastAPI and enmine rothal serief al
- Privacy-aware uploading to medical massiffiers
- Error-handling and procissure secrerary

EVALUATION & METRICS

Metric	Why needed	Recommendation
Accuracy	Recommended accuracy	Accuracy
Macro-F1	Micro-Flobal encoding	Macro-F1
Sensitivity/Specificity	Sensitivity/specificity	Sensitivity
ROC-AUC	Recommn ROC-AUC	ROC-AUC
ECE	Latency encerate ECE	ECE
Latency	Latency on row + latency	Latency

EVALUATION & METRICS

Metric	Why needed	Recommendation
Accuracy	Accuracy accuracy	Accuracy
Macro-F1	Macro-F1 mesoing	Macro-F1
Sensitivity/Specificity	Sensitivity/specificity	High
ROC-AUC	ROC and ROC-AUC	High
ECE	Latency emerate ECE	ECE
Latency	Latency on row + latency	Latency

CONCLUSION

- The path from the root-tvərm review to review to a deployable, verifiable hybrid prototype
- The lighting path to a deployable, verification in monsistie, verifiation: tororand, verifiable hybrid prototype, and unlitmate conctent.

Virtual University of Pakistan | University of Engineering and Technology Lahore | Coventry University

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

DECLARATIONS

Author Contributions

Muhammad Mubeen Gorya: conceptualization, literature collection, proposed system design, software development [confirm after implementation], experimentation [insert after completion], data analysis [insert after completion], and manuscript preparation. Muhammad Kaleem Ullah: supervision and methodological guidance [confirm contribution wording before submission].

Funding

No external funding was reported for this work.

Data and Code Availability

No empirical dataset or code repository accompanied the source project document reviewed for this manuscript. The final submission should provide a repository link, model configuration, and dataset-access statement after implementation.

Conflict of Interest

The author declares no known conflict of interest.

Ethics Statement

This review-and-design manuscript does not report experiments involving human participants or identifiable clinical data. Any future use of patient-derived images must follow dataset licenses, institutional ethics requirements, de-identification procedures, and applicable data-protection rules.

Appendix A. High-Level Training and Inference Algorithm

1. Input: labeled training images D, pretrained CNN backbone B, transformer encoder T, fusion module G, classifier C.
2. Split D at subject level into training, validation, and test partitions.
3. For each training image: validate, decode, augment, resize, and normalize.
4. Extract multiscale CNN maps F1, F2, F3 = B(x).
5. Project and pool maps; convert them to tokens Z with scale and position embeddings.
6. Compute transformer representation t = T(Z) and pooled CNN representation c.
7. Fuse representations h = G(c, t); predict probabilities p = softmax(C(h)).
8. Optimize a prespecified loss; monitor validation macro-F1/AUC and stop early.
9. Calibrate the selected model on held-out validation predictions.
10. At inference: repeat deterministic preprocessing, calculate p, uncertainty, Grad-CAM, and attention rollout; return the output through the web API.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

References

Ahmad, I. S., Dai, J., Xie, Y., & Liang, X. (2025). Deep learning models for CT image classification: A comprehensive literature review. Quantitative Imaging in Medicine and Surgery, 15(1), 962–1011. https://doi.org/10.21037/qims-24-1400

Chen, C. (2025). A review of convolutional neural network based methods for medical image classification. Computers in Biology and Medicine, 170, 109507. https://doi.org/10.1016/j.compbiomed.2024.109507

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 6824–6835.

Gao, Z., Tian, Y., Lin, S. C., & Lin, J. (2025). A CT image classification network framework for lung tumors based on pre-trained MobileNetV2 model and transfer learning, and its application and market analysis in the medical field. arXiv. https://doi.org/10.48550/arXiv.2501.04996

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 1321–1330.

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.

Hu, J., Xiang, Y., Lin, Y., Du, J., Zhang, H., & Liu, H. (2025). Multi-scale transformer architecture for accurate medical image classification. Proceedings of the 2025 International Conference on Artificial Intelligence and Computer Engineering. https://doi.org/10.1145/3730436.3730505

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.

Jiwar, Z. M., & Othman, Z. I. (2023). Sidr tree disease diagnosis system Python programming. Iraqi Journal of Intelligent Computing and Informatics, 2(1), 26–35.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). MViTv2: Improved multiscale vision transformers for classification and detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4804–4814.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. International Conference on Learning Representations.

Peng, Y. (2025). Research on the application of deep learning methods in the field of image classification. ITM Web of Conferences, 70, 04036. https://doi.org/10.1051/itmconf/20257004036

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., &Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510–4520.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

Shabbir, A., Ali, A., Ahmed, N., Zafar, J., Rasheed, B., Sajid, A., Ahmed, M., Afzal, A., Saadat, D., & Hanif, H. (2021). Satellite and scene image classification based on transfer learning and fine tuning of ResNet50. Mathematical Problems in Engineering, 2021, 5843816. https://doi.org/10.1155/2021/5843816

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, 6105–6114.

Author Note

This manuscript is derived from the BSCS final-year project "Image Classification Web Application." The available project document contains a literature corpus and a proposed hybrid approach, but it does not contain a verifiable dataset, source code audit, training logs, or experimental results. Accordingly, this article is written as a review-and-design contribution and does not claim empirical performance.