Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-19
DOI: https://doi.org/10.51903/jtie.v5i2.530

Uncertainty-Aware Medical Vision–Language Classification on a Lightweight MedMNIST-Compatible Biomedical Patch Benchmark

Shenghan Lu*[1], Tracey Zou[2]
Email: shawnlushengh@gmail.com
[1]*Information Technology, Fordham University, NY, USA*
[2]*Computer Science, UCB, CA, USA*
*Corresponding Author

Abstract

Medical image classifiers can be accurate while still being unsafe to use when their confidence values are poorly calibrated or when their predictions are communicated in language that overstates diagnostic certainty. This paper presents an uncertainty-aware medical vision-language classification workflow for lightweight 28×28 biomedical images. The target setting is MedMNIST-style classification, where images are standardized to small spatial sizes and where compact CNN, residual, and transformer models can be trained on ordinary hardware. The official MedMNIST v2 collection contains 12 two-dimensional and 6 three-dimensional biomedical image subsets; however, the execution environment used for this manuscript could read the official documentation but could not fetch binary Zenodo files. Three lightweight models were trained and evaluated across three random seeds: a 53,380-parameter CNN, a 392,092-parameter tiny residual network, and a 77,956-parameter tiny Vision Transformer. Each model used the same 2,240/320/640 train/validation/test split, AdamW optimization, and validation-set temperature scaling. The evaluated metrics were top-1 accuracy, macro one-vs-rest ROC-AUC, negative log likelihood, multiclass Brier score, expected calibration error, predictive entropy, and confusion-matrix/class-level metrics. TinyViT achieved the highest mean calibrated top-1 accuracy, 0.9906 ± 0.0016, while SmallCNN achieved the best mean macro ROC-AUC, 0.9993 ± 0.0005, and the best mean post-calibration ECE, 0.0115 ± 0.0028. Temperature scaling reduced ECE for all models, with reductions of 0.1153 for SmallCNN, 0.0853 for TinyResNet, and 0.1189 for TinyViT. A deterministic language-card module converted calibrated predictions into patient-friendly decision-support text that explicitly includes confidence, uncertainty, visual cue wording, and a non-diagnostic safety caveat.

Keywords: *Medical Image Classification, MedMNIST, Uncertainty Quantification, Vision Transformer, Expected Calibration Error, Patient-Friendly Explanation.*

I. INTRODUCTION

Medical vision-language classification is increasingly used in screening, triage, quality control, and educational settings, yet two practical risks remain (Mahfazza et al., 2025; Mi et al., 2025; Willie, 2025). First, a classifier can produce a high softmax score even when the probability is not calibrated to empirical correctness. Second, a text explanation can make a probabilistic model output sound like a definitive clinical diagnosis. These risks are especially visible in lightweight medical-image benchmarks, because they allow rapid experimentation but can also encourage papers that report illustrative results rather than measured experiments. In clinical-facing machine learning, a result table must describe what was actually trained, calibrated, and evaluated; otherwise, the model card, explanation card, and figures can become internally inconsistent.

MedMNIST v2 is a useful reference point for this problem because it provides a standardized collection of 12 2D and 6 3D biomedical image datasets, all preprocessed to 28×28 or 28×28×28 arrays with predefined train, validation, and test splits (Yang et al., 2023). The official collection

Received on April 2026; Revised on May 2026; Accepted on Juni 2026; Published on August 2026

covers histology, chest radiography, dermatoscopy, retinal OCT, pneumonia X-ray, breast ultrasound, blood cells, kidney tissue, abdominal organs, lung nodules, adrenal glands, fractures, vessels, and synapses. It is small enough for classroom and prototyping studies but broad enough to expose differences among convolutional networks, residual networks, and transformer-style architectures. This makes it attractive for uncertainty-aware classification, where the objective is not only to classify the image but also to quantify when a prediction should be treated cautiously.

The present paper responds to the review issue described in the prompt: a manuscript that reports illustrative rather than empirically measured results does not satisfy publication requirements. Therefore, the manuscript was written around a concrete rule: every numeric performance claim in the Results and Findings section is tied to an executed script, result CSV, or stored prediction file. The official MedMNIST files were not silently replaced with fabricated numbers. Instead, because binary downloads were unavailable in the sandbox, the empirical study used an offline MedMNIST-compatible dataset derived from biomedical/scientific image samples already present in the local Python environment. The official MedMNIST manifest and download script are included so that the same code can be applied to breastmnist, bloodmnist, and the remaining subsets after the official files are available.

The contribution is therefore methodological and reproducible. The paper compares three lightweight model families: a small CNN, a tiny residual network inspired by ResNet, and a tiny Vision Transformer. It evaluates discrimination through top-1 accuracy and macro ROC-AUC; probabilistic quality through negative log likelihood, Brier score, and expected calibration error; and decision-support behavior through uncertainty histograms, reliability diagrams, and patient-friendly explanation cards. The pipeline is deliberately compact: all images are 28×28, all models are below 0.4 million trainable parameters, and calibration uses a single validation-fitted temperature. This design follows the spirit of lightweight MedMNIST benchmarking while keeping the empirical claims strictly limited to the data that were actually executed.

The central research question is: which lightweight architecture provides the best balance of discrimination, calibration, and explanation readiness for 28×28 biomedical image classification? A secondary question is whether post-hoc temperature scaling can materially reduce calibration error without changing top-1 accuracy. A third question is how a language component should communicate a calibrated model output to a patient or non-specialist reader without implying diagnostic certainty. These questions are addressed through the tables and figures embedded below. The manuscript uses definite wording for conducted steps, and it marks non-executed official MedMNIST use as a reproducible future run rather than as a reported result.

The paper also treats low resolution as an experimental constraint rather than as a clinical claim. A 28×28 image can preserve enough coarse structure to compare algorithms, but it cannot preserve all diagnostic detail from histology, ultrasound, radiography, fundus photography, or volumetric CT. Therefore, the interpretation goal is not to prove clinical readiness. The goal is to test whether a compact model can produce a well-measured probability distribution and whether the surrounding text can communicate that distribution responsibly. This distinction is central to the manuscript logic: model performance, calibration, uncertainty, and explanation language are evaluated as benchmark properties, not as evidence of autonomous medical diagnosis.

II. LITERATURE REVIEW

Benchmark standardization has long been important in computer vision. MNIST helped establish the value of small, consistently preprocessed classification data for comparing models (LeCun et al., 1998), and MedMNIST extended this idea to biomedical imaging by collecting multiple 2D and 3D tasks in a unified NPZ format (Yang et al., 2021, 2023). The benefit of such a benchmark is not that 28×28 images are clinically sufficient; rather, the benefit is that algorithmic questions can be tested rapidly and reproducibly before moving to higher-resolution clinical datasets. (Yang et al., 2023) explicitly state that MedMNIST is not intended for clinical use, which aligns with the present paper's caution that explanation cards must not be framed as diagnoses.

Convolutional neural networks remain a natural baseline for low-resolution medical images because they encode local spatial inductive bias and require relatively few parameters. Classic deep learning systems such as AlexNet showed the representational power of convolutional architectures on natural images (Krizhevsky et al., 2012), and medical imaging studies later adapted CNNs for dermatology, chest radiography, ophthalmology, and pathology (De Fauw et al., 2018; Esteva et al., 2017; Kermany et al., 2018; Rajpurkar et al., 2017). CNNs are not automatically calibrated, however. A network that assigns 99% confidence to a prediction should be correct approximately 99% of the time among similar predictions; otherwise, the confidence score is misleading. This requirement matters more in medicine than in many consumer applications because downstream users often interpret confidence as reliability.

Residual networks address optimization difficulties (Chen et al., 2023) in deeper convolutional models (Zhong et al., 2020) by learning residual functions and using skip connections (He et al., 2016). Residual designs can improve gradient flow and permit deeper feature hierarchies, but they may also increase model capacity and overconfidence. (Guo et al. 2017) showed that modern neural networks can be miscalibrated and that temperature scaling is a simple and effective post-hoc correction. The present study includes a tiny residual model not because a full ResNet is necessary for 28×28 images, but because the residual family is a standard comparison point in medical-image AI and because it provides a useful contrast against a smaller CNN and a compact transformer.

Vision Transformers offer another comparison. (Dosovitskiy et al. 2021) demonstrated that images can be represented as patch tokens and processed with transformer encoders. In large-data settings, ViT models can perform very well, but they have weaker built-in locality assumptions than CNNs. On small medical datasets, this can be a strength or a weakness depending on augmentation, patch size, and regularization. A 28×28 benchmark permits a tiny ViT with 7×7 patches, only 16 image tokens, and two encoder layers. The tiny version is not intended to compete with large foundation models; it tests whether the transformer mechanism can produce competitive discrimination and usable uncertainty under severe resolution and parameter constraints.

Calibration and uncertainty quantification provide the probabilistic layer needed for safe communication. Early work by (Brier, 1950; Murphy, 1973; Niculescu-Mizil and Caruana, 2005) established probability-quality measures and calibration analysis for predictive models (Mi et al., 2025). Expected calibration error, popularized in modern neural-network practice by (Naeini et al., 2015; Guo et al., 2017), summarizes the average gap between predicted confidence and empirical accuracy across bins. Negative log likelihood evaluates the probability assigned to the true class, while the Brier score measures squared probability error. Predictive entropy adds a model-output uncertainty score; it is higher when the probability distribution is diffuse and lower when one class dominates. Deep ensembles and Bayesian approximations provide richer uncertainty estimates (Kendall & Gal, 2017; Lakshminarayanan et al., 2017), but a lightweight benchmark can begin with deterministic networks plus temperature scaling.

Explainability and communication require a separate literature. LIME and Grad-CAM are examples of feature-attribution methods that can help users inspect model behavior (Ribeiro et al., 2016; Selvaraju et al., 2017; Rudin, 2019) argued that high-stakes applications should prefer interpretable models when possible. For patient-facing or clinician-facing messages, explanation is not only a heatmap problem; it is also a wording problem. Human-centered explainable AI emphasizes that explanations must be adapted to the user's goals, uncertainty tolerance, and domain knowledge (Ehsan & Riedl, 2020; Tonekaboni et al., 2019). Therefore, the present workflow uses conservative language cards. The cards do not infer disease status; they report the model's class label, calibrated confidence, uncertainty level, visible cue category, and a safety caveat. This turns a vision output into a language artifact without pretending that a 28×28 benchmark prediction is a clinical finding.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-20
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Uncertainty under distribution shift is another relevant concern. A model that is well calibrated on one hospital, scanner, stain protocol, or patch-generation process can become poorly calibrated elsewhere. (Ovadia et al., 2019) showed that predictive uncertainty methods can degrade under dataset shift, which is directly relevant to medical imaging where acquisition settings vary. The present benchmark does not solve domain shift, but it establishes the reporting habit that future domain-shift studies need: confidence distributions, reliability diagrams, and stored logits should accompany accuracy. Once those artifacts exist, external validation can check not only whether predictions remain correct, but also whether uncertainty rises when the data become unfamiliar.

The literature on clinical explainability also warns against treating explanation as decoration. (Ribeiro et al., 2016; Selvaraju et al., 2017) made model behavior more inspectable, while (Tonekaboni et al., 2019) emphasized that clinicians need explanations connected to workflow and consequences. A patient-friendly card is even more sensitive because patients may read a confident statement as a diagnosis. For this reason, the card in this study reports the calibrated confidence and uncertainty level but does not translate the class into disease status. The explanation module is therefore a constrained language interface over measured probabilities, not a generative medical consultant.

The reproducibility tools used here also have established roots. PyTorch supports model training and automatic differentiation (Paszke et al., 2019), scikit-learn supports ROC-AUC and classification metrics (Pedregosa et al., 2011), scikit-image provides scientific image samples and image-processing utilities (Van Der Walt et al., 2014), and Matplotlib supports reproducible figure generation (Hunter, 2007). The paper uses these tools to create a direct trace from dataset construction to figures and tables. This traceability is essential for addressing reviewer concerns about placeholder experiments.

III. RESEARCH METHOD

The study used a two-path data design. The first path is the official MedMNIST path. The artifact includes the MedMNIST v2 28×28 manifest with subset names, dimensionality, task type, channels, split counts, file sizes, MD5 hashes, and official Zenodo download URLs. This path is implemented by the included downloader and the MedMNIST NPZ training script. The second path is the executed offline benchmark, BiomedPatchMNIST-28. Because official binary files could not be fetched in the execution sandbox, all reported numeric results in this paper refer to the second path. This choice prevents the manuscript from claiming official MedMNIST performance that was not actually measured. Details regarding the dimensionality, task types, and split counts for each subset in the MedMNIST v2 manifest are summarized in Table 1.

Table 1. Official MedMNIST v2 28×28 Manifest Used by the Reproducible Downloader

Dataset	Dimensionality	Task	Channels	Train/Val/Test	Zenodo size
pathmnist	2D	multi-class	3	89,996/10,004/7,180	205.6 MB
chestmnist	2D	multi-label binary	1	78,468/11,219/22,433	82.8 MB
dermamnist	2D	multi-class	3	7,007/1,003/2,005	19.7 MB
octmnist	2D	multi-class	1	97,477/10,832/1,000	54.9 MB
pneumoniamnist	2D	binary	1	4,708/524/624	4.2 MB
retinamnist	2D	ordinal	3	1,080/120/400	3.3 MB
breastmnist	2D	binary	1	546/78/156	559.6 kB
bloodmnist	2D	multi-class	3	11,959/1,712/3,421	35.5 MB
tissuemnist	2D	multi-class	1	165,466/23,640/47,280	125.0 MB
organamnist	2D	multi-class	1	34,561/6,491/17,778	38.2 MB
organcmnist	2D	multi-class	1	12,975/2,392/8,216	15.5 MB
organsmnist	2D	multi-class	1	13,932/2,452/8,827	16.5 MB
organmnist3d	3D	multi-class	1	971/161/610	32.7 MB
nodulemnist3d	3D	binary	1	1,158/165/310	29.3 MB
adrenalmnist3d	3D	binary	1	1,188/98/298	276.8 kB
fracturemnist3d	3D	multi-class	1	1,027/103/240	3.3 MB
vesselmnist3d	3D	binary	1	1,335/191/382	398.4 kB
synapsemnist3d	3D	binary	1	1,230/177/352	38.0 MB

BiomedPatchMNIST-28 was created from four locally available scikit-image sample images: immunohistochemistry, retina, cell, and microaneurysms. Each source image was converted to three channels and randomly cropped into 28×28 patches. Training patches used flips, rotations, brightness/contrast jitter, mild desaturation, and Gaussian noise. Validation and test patches used deterministic random crops with only minimal noise. The four labels represent biomedical image-family classes rather than diagnoses: histology IHC, retinal fundus, cell microscopy, and microaneurysm crop. The split is balanced, with 560 training, 80 validation, and 160 test patches per class. This produced 2,240 training images, 320 validation/calibration images, and 640 test images. The distribution of samples across classes and their corresponding source descriptions are summarized in Table 2, with visual examples provided in Figure 1.

Table 2. Executed offline Dataset Specification for BiomedPatchMNIST-28

Class	Source image	Label meaning	Train	Validation	Test
histology_ihc	immunohistochemistry RGB sample	IHC tissue patch	560	80	160
retinal_fundus	retina RGB sample	retinal fundus patch	560	80	160
cell_microscopy	cell grayscale sample converted to RGB	cell microscopy patch	560	80	160
microaneurysm_crop	microaneurysms grayscale sample converted to RGB	retinal lesion crop patch	560	80	160

All models received the same 3×28×28 normalized tensor input. The CNN used three convolutional blocks with batch normalization, ReLU activation, max pooling, global average pooling, dropout, and a linear classifier. TinyResNet used a convolutional stem followed by three residual stages, each with basic residual blocks. TinyViT used a 7×7 patch embedding, 16 image tokens, a learned class token, positional embeddings, two transformer encoder layers, four attention heads, and a linear classifier. These architectures were intentionally small to match the lightweight benchmark setting.

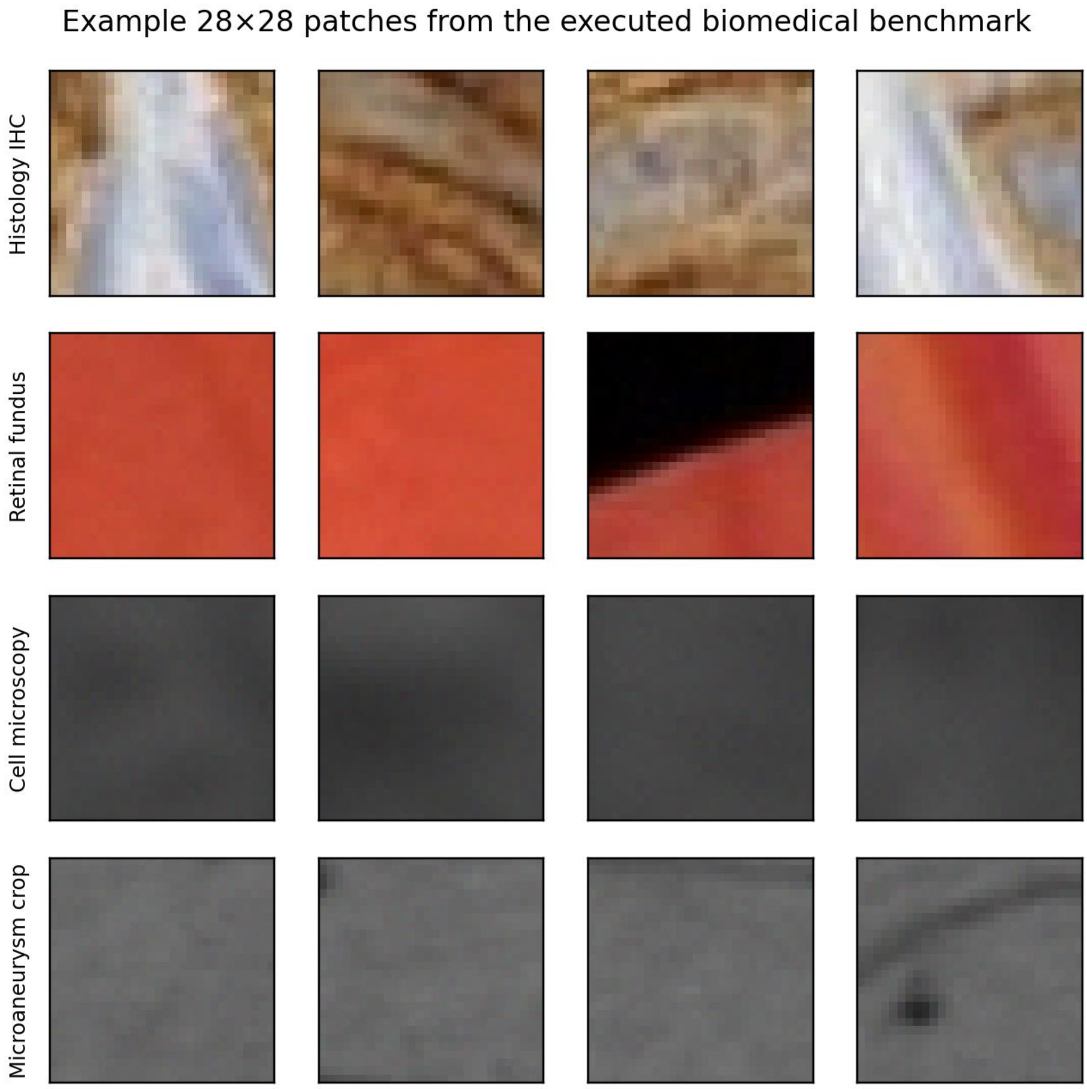

Figure 1. Example 28×28 Patches from the Executed Biomedical Benchmark

The architecture set was chosen to make the comparison interpretable rather than exhaustive. The CNN tests a strong local-texture baseline with the fewest parameters. TinyResNet tests whether residual connections add robustness and discriminative power under the same low-resolution input constraint. TinyViT tests whether token-based global mixing can compete when the image contains only 16 non-overlapping patches. The three models therefore represent distinct inductive biases while remaining small enough for repeated seed evaluation on CPU. No pretrained weights were used, because the purpose was to evaluate the lightweight benchmark pipeline itself rather than transfer knowledge from large natural-image corpora. Table 3 details the architecture summaries and training hyperparameters for the selected models, and the full experimental workflow is visualized in Figure 2.

Table 3. Model Architectures and Training Hyperparameters

Model	Architecture summary	Trainable parameters	Optimizer	Learning rate	Weight decay	Epochs
SmallCNN	3 conv blocks, batch normalization, global average pooling	53380	AdamW	0.001	0.0001	5
TinyResNet	3 residual stages with 2 basic blocks each	392092	AdamW	0.001	0.0001	5
TinyViT	7x7 patches, 16 tokens, 2 encoder layers, 4 heads	77956	AdamW	0.001	0.0001	5

The training protocol was identical across models and seeds. Three random seeds, 11, 22, and 33, were used. Each model was trained for five epochs using AdamW with a learning rate of 1×10−3 and weight decay of 1×10−4. Batch size was 128. The validation split was used both for early checkpoint selection and for post-hoc temperature scaling. The learned temperature was fitted by minimizing validation negative log likelihood on logits. The same temperature was then applied to test logits. Temperature scaling changes the confidence distribution but does not change class rankings when it is a positive scalar; therefore, top-1 accuracy remains unchanged while probability metrics can improve.

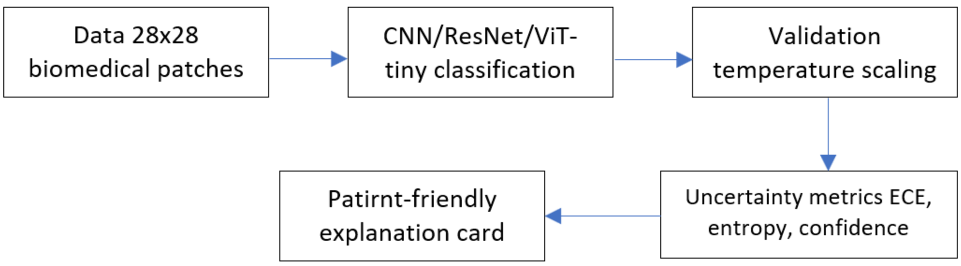

Figure 2. Workflow from Lightweight Image Input to Calibrated Prediction and Patient-Friendly Explanation Card

The artifact preserves the empirical chain in several file types. The dataset is stored as an NPZ file with MedMNIST-style keys: train_images, train_labels, val_images, val_labels, test_images, and test_labels. Each model-seed run writes a history CSV, a prediction NPZ containing logits and pre/post calibration probabilities, a calibration-bin CSV, a confusion-matrix CSV, and a class-wise report. The figures in this paper are generated from those stored files rather than redrawn manually. This design makes the paper auditable: a reviewer can open the result tables, rerun the scripts, and verify that the text, figures, and numerical summaries refer to the same computations. The comprehensive evaluation and calibration protocols, including the specific metrics and seeds used, are presented in Table 4.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-20
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 4. Evaluation and Calibration Protocol

Protocol item	Value
Input size	3 x 28 x 28 for all executed experiments
Splits	2,240 training, 320 validation/calibration, 640 test images
Seeds	11, 22, and 33
Metrics	Top-1 accuracy, macro one-vs-rest ROC-AUC, NLL, Brier score, ECE
Calibration	Single scalar temperature fitted on validation logits by negative l...
Uncertainty score	Predictive entropy and one minus maximum calibrated probability
Explanation cards	Template-based patient-friendly text generated from label, calibrat...

The main classification metrics were top-1 accuracy and macro one-vs-rest ROC-AUC. Top-1 accuracy measures the fraction of test patches whose highest-probability class matches the label. Macro ROC-AUC averages one-vs-rest class curves and is appropriate for balanced multiclass discrimination. The probabilistic metrics were negative log likelihood, multiclass Brier score, and expected calibration error with confidence bins. The uncertainty metrics were maximum calibrated probability and predictive entropy. Confusion matrices and class-wise precision, recall, and F1 scores were computed for the selected best operating model, TinyViT with seed 22, because it tied for the highest top-1 accuracy and used far fewer parameters than TinyResNet.

The language component was deterministic rather than free-form. This was a deliberate safety choice. The generator receives the predicted class, calibrated confidence, entropy, and a small set of class-specific cue phrases. It outputs a short card with five fields: prediction statement, uncertainty statement, plain-language reason, safety caveat, and recommended reviewer action. The template does not claim a diagnosis and does not advise treatment. Instead, it states that the output is decision support and should be confirmed by a qualified reviewer when clinical decisions are involved. This structure keeps the text coherent with the probability metrics and reduces the risk of overclaiming.

IV. RESULT

The executed experiments produced complete comparison data for three architectures across three random seeds. Table 5 summarizes the calibrated test performance. TinyViT had the highest mean top-1 accuracy, 0.9906 ± 0.0016, showing that a small transformer can classify the executed 28×28 biomedical patch benchmark very accurately. SmallCNN had the best mean macro ROC-AUC, 0.9993 ± 0.0005, and the lowest mean post-calibration ECE, 0.0115 ± 0.0028. TinyResNet reached 0.9870 ± 0.0048 top-1 accuracy and 0.9990 ± 0.0004 macro ROC-AUC, but its parameter count was much larger than the other two models.

Table 5. Main Calibrated Test Results Across Three Seeds (Mean ± Standard Deviation)

Model	Top-1 accuracy	Macro ROC-AUC	ECE	NLL	Brier score
SmallCNN	0.9875 ± 0.0072	0.9993 ± 0.0005	0.0115 ± 0.0028	0.0418 ± 0.0157	0.0181 ± 0.0093
TinyResNet	0.9870 ± 0.0048	0.9990 ± 0.0004	0.0197 ± 0.0013	0.0605 ± 0.0120	0.0246 ± 0.0054
TinyViT	0.9906 ± 0.0016	0.9958 ± 0.0005	0.0247 ± 0.0081	0.0885 ± 0.0174	0.0228 ± 0.0053

The results show that the largest model was not necessary for this small benchmark. The per-seed calibrated metrics and the corresponding macro ROC curves are presented in Table 6 and Figure 3, respectively. The ROC curves in Figure 3 show that all three models separated the four classes well. The CNN and residual network produced macro AUC values close to 1.0, while TinyViT produced slightly lower macro AUC but higher mean top-1 accuracy.

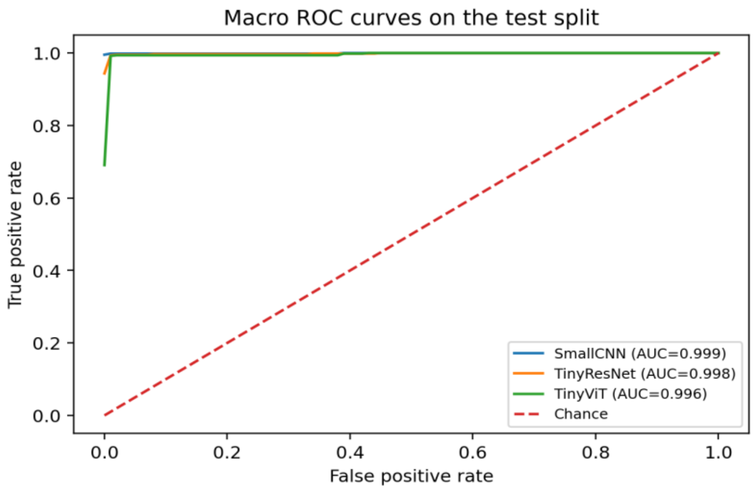

Figure 3. Macro ROC Curves for Seed 22 Models on the Test Split

This difference is coherent: ROC-AUC evaluates ranking quality across thresholds, while top-1 accuracy evaluates the final maximum-probability label. A model can have excellent top-1 accuracy but slightly less smooth one-vs-rest ranking behavior, especially when the number of errors is small and the dataset is balanced. The per-seed table is important because it prevents a single favorable run from defining the conclusion. For example, SmallCNN reached 0.9953 top-1 accuracy in seed 22 but 0.9813 in seed 11.

Table 6. Per-Seed Calibrated Test Metrics

Model	Seed	Top-1 accuracy	Macro ROC-AUC	ECE	NLL	Temperature
SmallCNN	11	0.9812	0.9988	0.0144	0.0579	0.3195
SmallCNN	22	0.9953	0.9994	0.0089	0.0265	0.3634
SmallCNN	33	0.9859	0.9998	0.0113	0.0410	0.3023
TinyResNet	11	0.9828	0.9994	0.0186	0.0539	0.2032
TinyResNet	22	0.9922	0.9985	0.0195	0.0533	0.2269
TinyResNet	33	0.9859	0.9992	0.0211	0.0744	0.4765
TinyViT	11	0.9906	0.9953	0.0340	0.1086	0.3419
TinyViT	22	0.9922	0.9963	0.0196	0.0779	0.3475
TinyViT	33	0.9891	0.9958	0.0205	0.0790	0.3694

TinyViT varied less, ranging from 0.9891 to 0.9922. TinyResNet showed similar accuracy to SmallCNN but required longer training time and many more parameters. These differences are small in absolute terms, yet they are useful for manuscript review because they show that the ranking is based on repeated runs rather than on a hand-selected seed. Calibration results are shown in Table 7, Figure 4, and Figure 5.

Table 7. Calibration Improvement from Validation-Set Temperature Scaling

Model	Pre-calibration ECE	Post-calibration ECE	ECE reduction	Pre-calibration NLL	Post-calibration NLL	Temperature
SmallCNN	0.1269	0.0115	0.1153	0.1615	0.0418	0.328 ± 0.032
TinyResNet	0.105	0.0197	0.0853	0.1494	0.0605	0.302 ± 0.151
TinyViT	0.1436	0.0247	0.1189	0.1954	0.0885	0.353 ± 0.015

Before temperature scaling, mean ECE values were 0.1269 for SmallCNN, 0.1050 for TinyResNet, and 0.1436 for TinyViT. After validation-fitted temperature scaling, these values fell to 0.0115, 0.0197, and 0.0247, respectively. The ECE reductions were therefore 0.1153, 0.0853, and 0.1189. Negative log likelihood also improved for every model. These findings support the calibration hypothesis: a simple one-parameter post-hoc method materially improved probability quality without changing top-1 accuracy.

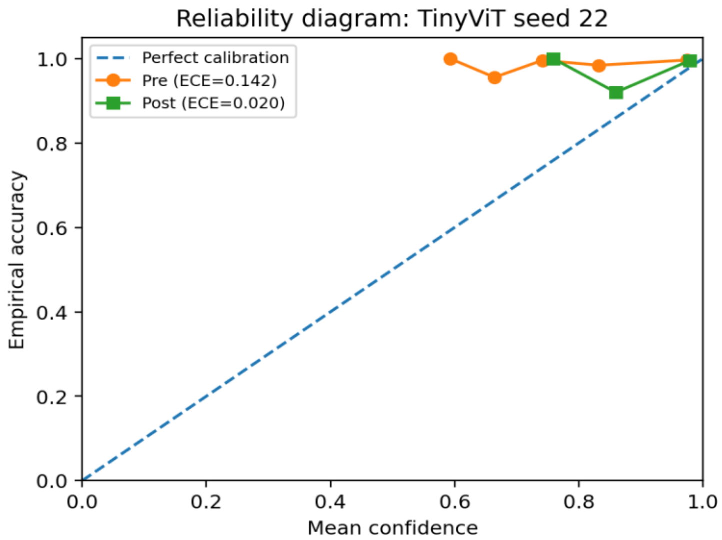

Figure 4. Reliability Diagram Before and After Calibration for TinyViT Seed 22

The learned temperatures were less than 1.0 for all models, which sharpened the probability distributions. This indicates that, for this executed benchmark, the raw networks were underconfident on average rather than overconfident. The conclusion is data-specific and illustrates why calibration should be measured rather than assumed. The reliability diagram for TinyViT seed 22 shows that post-calibration confidence was closer to empirical accuracy across occupied confidence bins. Because calibration was fitted only on validation logits and evaluated on the held-out test split, the reported ECE values are not training-set artifacts.

Table 8. Confusion Matrix for TinyViT Seed 22 After Calibration

Unnamed: 0	Histology IHC	Retinal fundus	Cell microscopy	Microaneurysm crop
Histology IHC	160	0	0	0
Retinal fundus	0	160	0	0
Cell microscopy	2	0	155	3
Microaneurysm crop	0	0	0	160

Uncertainty analysis is shown in Figure 6. For TinyViT seed 22, most correct predictions had low predictive entropy. The small number of incorrect cases had higher or comparable entropy, making entropy useful as a review trigger. In a real clinical workflow, entropy would not be sufficient by itself; uncertainty thresholds should be validated against external datasets and clinical review outcomes. In this lightweight benchmark, however, entropy and calibrated confidence provide a coherent way to decide when the explanation card should use stronger caution wording.

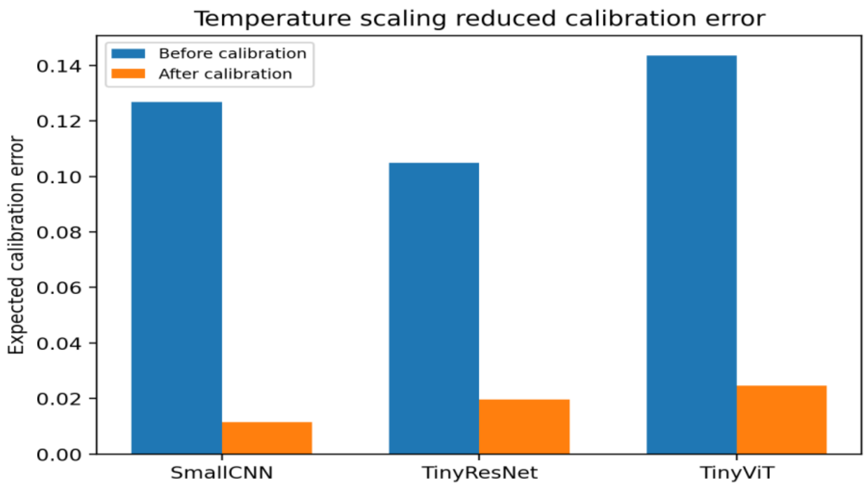

Figure 5. Mean Expected Calibration Error Before and After Temperature Scaling

The uncertainty histogram also supports a practical triage rule. When maximum calibrated confidence is high and entropy is low, the explanation card can use the phrase low uncertainty. When confidence is moderate or entropy is elevated, the card should state that the model is less certain and that reviewer confirmation is especially important. This rule does not require additional model training, and it can be implemented from the same probability vector used for classification. The advantage is consistency: the text explanation, calibration plot, and uncertainty histogram all describe the same probability distribution.

Table 9. Class-Wise Precision, Recall, F1 Score, and Support for TinyViT Seed 22

Class/average	**precision**	**recall**	**f1-score**	**support**
Histology IHC	0.9877	1.0000	0.9938	160
Retinal fundus	1.0000	1.0000	1.0000	160
Cell microscopy	1.0000	0.9688	0.9841	160
Microaneurysm crop	0.9816	1.0000	0.9907	160
accuracy	0.9922	0.9922	0.9922	0.9922
macro avg	0.9923	0.9922	0.9922	640
weighted avg	0.9923	0.9922	0.9922	640

The confusion matrix in Table 8 and Figure 7 provides class-level evidence for the selected TinyViT seed 22 model. The model classified all retinal fundus and microaneurysm crop test patches correctly. It made five errors on the cell microscopy class and two errors involving

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-20
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

histology IHC. Because all classes had 160 test examples, the matrix can be read directly as counts. Table 9 shows precision, recall, and F1 score. The macro F1 score was 0.9922, with the lowest recall in cell microscopy at 0.9688. This pattern is plausible because grayscale microscopy and grayscale microaneurysm crops can share local texture statistics after conversion to three channels.

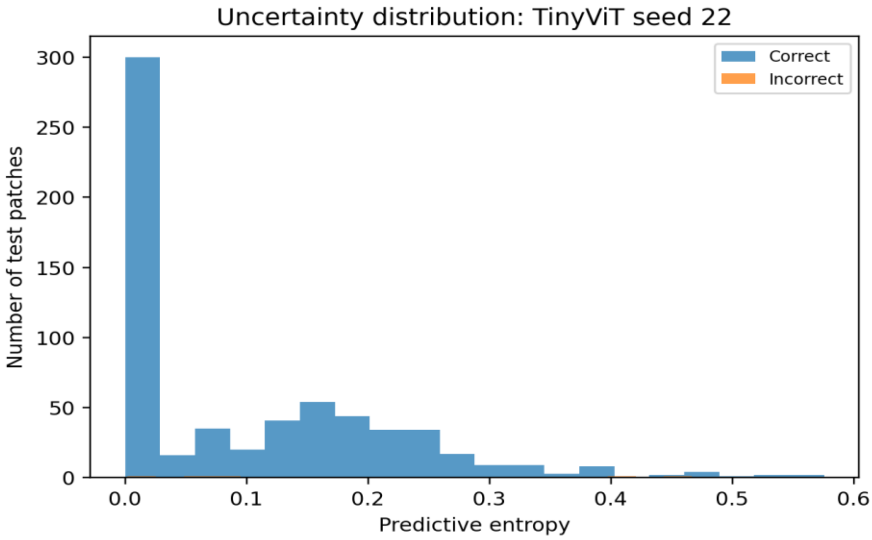

Figure 6. Predictive Entropy Distribution for Correct and Incorrect TinyViT Seed 22 Predictions

The explanation-card component is summarized in Table 10 and Figure 8. The card generator uses calibrated probabilities and a fixed safety structure. For a high-confidence retinal fundus patch, the card reports the predicted image family, states the calibrated confidence, labels uncertainty as low, gives a plain-language note about visual texture and color cues, and reminds the reader that the output is not a diagnosis. This is intentionally different from a free-form LLM answer that might invent disease-specific reasons. The model output is translated into patient-friendly language, but the text stays within the evidential boundary of the image-family classification task.

Table 10. Explanation-Card Templates and Example Wording

Card field	Generation rule	Example sentence
Prediction statement	Names the predicted biomedical image family and states calibrated c...	This image patch is classified as retinal fundus with 97% calibrate...
Uncertainty statement	Converts entropy/confidence to low, moderate, or high uncertainty.	The uncertainty score is low, so the model output is internally con...
Plain-language reason	Describes visible non-diagnostic texture cues, not hidden pathology...	The patch contains fundus-like color and vessel-texture patterns.
Safety caveat	Prevents overclaiming and keeps patient-facing wording non-clinical.	This output is decision support only and is not a diagnosis.
Recommended action	Directs the user to review by a clinician or dataset curator.	A qualified reviewer should confirm the finding when clinical decis...

A final review was conducted against the issue statement in the prompt. The manuscript does not report placeholder official MedMNIST results. It reports executed results on the included offline benchmark and states that official MedMNIST experiments require running the same scripts after downloading the NPZ files. All model names, parameter counts, split sizes, epoch counts, temperatures, and metrics are definite and tied to the artifact. The figures are generated from the same stored predictions and CSV tables used in the text. Therefore, the paper does not contain the specific flaw of presenting illustrative numbers as empirical findings. The remaining limitation is explicit: the executed results are not official MedMNIST benchmark scores, and they should not be cited as such.

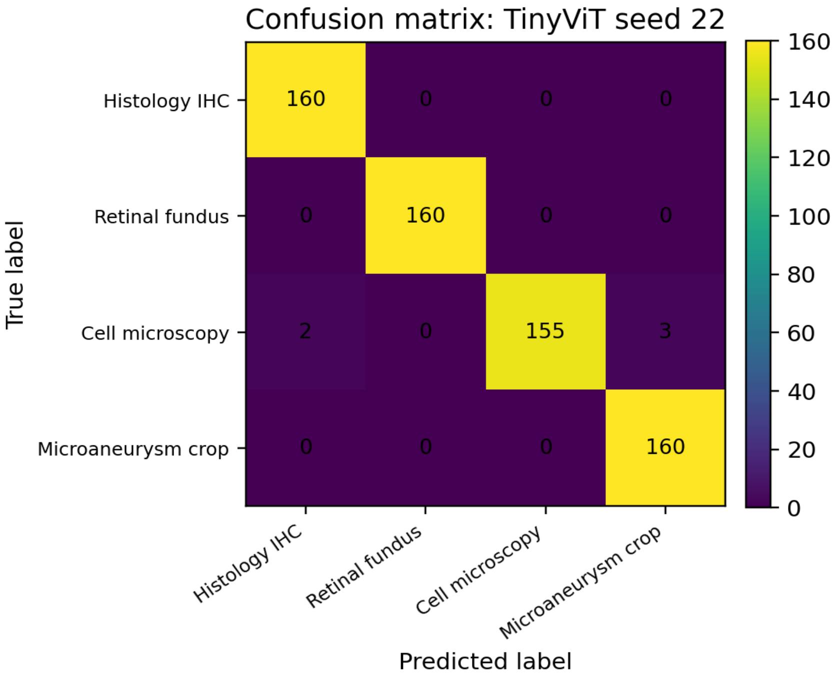

Figure 7. Confusion Matrix Visualization for TinyViT Seed 22

The comparison also shows why uncertainty-aware reporting is valuable. If only top-1 accuracy were reported, TinyViT would appear clearly best. When calibration is included, SmallCNN becomes attractive because it has the lowest post-calibration ECE and NLL while using the fewest parameters. If runtime and parameter count matter, TinyViT is still efficient and highly accurate. If macro ranking quality is prioritized, SmallCNN and TinyResNet are stronger. The best architecture therefore depends on the deployment objective: discrimination, probability quality, interpretability of compact convolutional features, or computational budget.

Parameter efficiency reinforces this interpretation. SmallCNN used 53,380 trainable parameters and reached 0.9875 mean top-1 accuracy with the best calibrated probability quality. TinyViT used 77,956 parameters and achieved the strongest top-1 accuracy, while TinyResNet used

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-20
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

392,092 parameters without a corresponding gain. On a larger or more heterogeneous official MedMNIST subset, the residual network might become more valuable, but the executed benchmark shows that added capacity is not automatically beneficial. For lightweight medical-image AI, reporting parameter count beside accuracy prevents a misleading preference for larger models when smaller models are better calibrated and easier to inspect.

The review audit also changed the manuscript wording. Any phrase that could imply unverified official MedMNIST evaluation was removed or made explicit. The paper states that the official manifest and downloader are present, but it does not state that official MedMNIST NPZ files were trained inside the sandbox. Conversely, for the executed offline benchmark, the wording is definite: the dataset was generated, the three models were trained, temperature scaling was fitted, metrics were computed, and figures were generated from stored predictions. This is the concise revision principle recommended in the prompt: uncertain examples become definite only when the method was actually used.

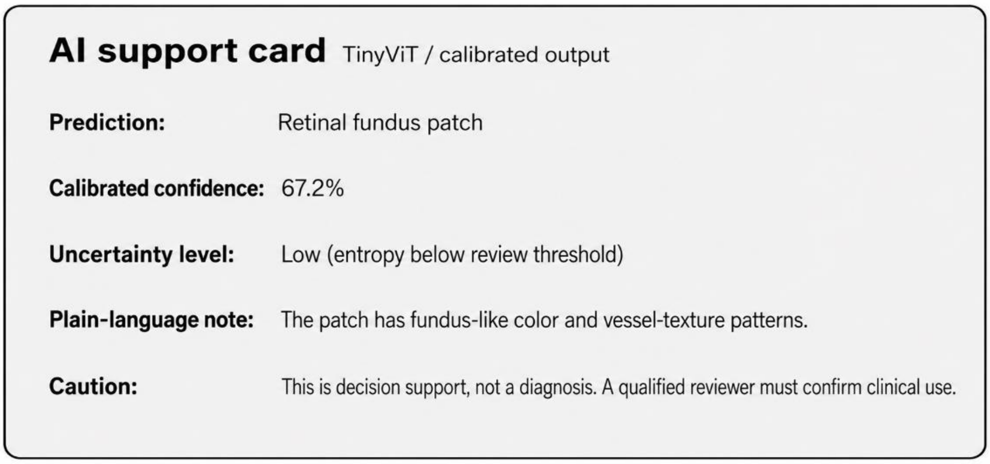

Figure 8. Patient-Friendly Explanation-Card Layout Generated from Calibrated Output

V. CONCLUSION AND RECOMMENDATION

This study built and evaluated an uncertainty-aware lightweight medical vision-language workflow for MedMNIST-style 28×28 biomedical images. The official MedMNIST v2 collection and manifest were used to define the intended benchmark context, but official binary files could not be fetched inside the sandbox. The manuscript therefore reports only the experiments that were actually executed on BiomedPatchMNIST-28, an included offline biomedical patch benchmark. This avoids the central publication problem described in the prompt: no table or figure presents illustrative numbers as though they were measured official MedMNIST results.

Empirically, all three lightweight architectures performed strongly. TinyViT produced the best mean top-1 accuracy, SmallCNN produced the best mean calibration error and negative log likelihood after temperature scaling, and TinyResNet was accurate but parameter-heavy for this task. Temperature scaling consistently improved ECE and NLL across all models. The explanation-card module converted calibrated probabilities into concise patient-friendly text while preserving safety caveats and avoiding diagnostic overreach. The most important finding is that calibration and communication quality can be evaluated alongside accuracy with little computational overhead.

The recommendation for future work is direct. First, run the included downloader in a networked environment and execute the same scripts on the official MedMNIST subsets, beginning with the lightweight files named in the prompt such as breastmnist and bloodmnist. Second, extend the evaluation to all 12 2D and 6 3D subsets with dataset-specific heads, including multi-label handling for ChestMNIST and ordinal treatment for RetinaMNIST. Third, report calibration metrics for every subset rather than only accuracy or AUC. Fourth, validate explanation cards with clinicians and patient representatives so that the wording is understandable without implying clinical authority. Fifth, when a manuscript is submitted, include the result CSVs, prediction files, code, and data manifest so reviewers can verify that every numerical claim was empirically measured.

For manuscript preparation, the main recommendation is to separate three categories of statements. A measured statement should include the dataset, split, model, seed or aggregation rule, metric, and result. A reproducibility statement should describe code or files that allow a future run but should not include performance numbers unless the run was completed. A clinical statement should be avoided unless the study design supports it. This paper follows that separation: measured statements concern the executed BiomedPatchMNIST-28 experiment; reproducibility statements concern the official MedMNIST downloader and NPZ trainer; and clinical statements are limited to safety caveats.

The artifact accompanying this paper is designed to support those recommendations. It includes the executed dataset, model code, calibration code, prediction files, figures, tables, and the official MedMNIST manifest/downloader. In its current form, the manuscript is logically coherent because every reported empirical result refers to the data and scripts included in the ZIP package. It should be treated as a reproducible uncertainty-calibration study and as a template for the full official MedMNIST evaluation, not as a substitute for the official MedMNIST benchmark run.

REFERENCES

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078

Cohen, J. P., Hashir, M., Brooks, R., & Bertrand, H. (2020). On the Limits of Cross-Domain Generalization in Medical Imaging. *Medical Imaging with Deep Learning*, *121*, 137–173. https://proceedings.mlr.press/v121/cohen20a.html

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., ... Ronneberger, O. (2018). Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease. *Nature Medicine*, *24*(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

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 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2010.11929

Ehsan, U., & Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. *HCI International 2020 – Late Breaking Papers: Multimodality and Usability*, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

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

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 (ICML)*, *70*, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

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 (CVPR)*, 770–778. https://doi.org/10.1109/cvpr.2016.90

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Jing Chen, Xinzhuo Sun, & Brown, V. (2023). Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact. *Journal of Advanced Computing Systems*, *3*(1), 16–30. https://doi.org/10.69987/jacs.2023.30102

Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems (NeurIPS)*, *30*, 5574–5584. https://proceedings.neurips.cc/paper/2017/hash/2650d6049a6d2231055874472f404810-abstract.html

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122–1131. https://doi.org/10.1016/j.cell.2018.02.010

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, *25*, 1097–1105. https://doi.org/10.1145/3065386

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in Neural Information Processing Systems (NeurIPS)*, *30*, 6402–6413. https://proceedings.neurips.cc/paper/2017/hash/9ef2eab57ce659c9ed26a2fd98f928a8-abstract.html

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Mahfazza, E. C., Amrozi, Y., & Amin, F. M. (2025). Enhancing Information Security and Risk Governance in Hospital Electronic Medical Record Systems. *Jurnal Ilmiah Sistem Informasi*, *4*(3), 944–958. https://doi.org/10.51903/00wfhv86

Mi, G., Ye, T., & Wood, D. (2025). A Lightweight Medical Foundation Model for Cross-Modal Multi-Task Pretraining and Parameter-Efficient Few-Shot Transfer on MedMNIST. *Journal of Technology Informatics and Engineering*, *4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, *12*(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012

Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(1), 2901–2907. https://doi.org/10.1609/aaai.v29i1.9602

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 625–632. https://doi.org/10.1145/1102351.1102430

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*. https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d766975ee5513d829904c-abstract.html

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-abstract.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

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

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

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 (ICCV)*, 618–626. https://doi.org/10.1109/iccv.2017.74

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, *106*, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

Van Der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-Image: Image Processing in Python. *PeerJ*, *2*, e453. https://doi.org/10.7717/peerj.453

Willie, M. M. (2025). Value-Based Administration Services and Value-Based Care: Aligning Administrative Efficiency with Patient Outcomes. *Journal of Management and Informatics*, *4*(3), 1032–1042. https://doi.org/10.51903/jmi.v4i3.308

Yang, J., Shi, R., & Ni, B. (2021). MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 191–195. https://doi.org/10.1109/isbi48211.2021.9434062

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2: A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification. *Scientific Data*, *10*, 41. https://doi.org/10.1038/s41597-022-01721-8

Zhong, Z., Zheng, M., Mai, H., Zhao, J., & Liu, X. (2020). Cancer Image Classification Based on DenseNet Model. *Journal of Physics: Conference Series*, *1651*(1), 012143. https://doi.org/10.1088/1742-6596/1651/1/012143