Indo American Journal of Multidisciplinary Research and Review (IAJMRR)
International Peer Reviewed - Refereed Research Journal
ISSN: 2581 - 6292, Impact Factor: 6.885, Website: www.iajmrr.com

# MODEL DEPTH, INPUT RESOLUTION, AND TRAINING DATASET SIZE AS PREDICTORS OF MEDICAL IMAGE CLASSIFICATION PERFORMANCE

**Divya Vishwanath* & Paul Johnson Asamoah****
* Janardhan Rai Nagar Rajasthan Vidyapeeth, Udaipur, Rajasthan, India
** Power Life University College, Accra, Ghana, West Africa

**Cite This Article:** Divya Vishwanath & Paul Johnson Asamoah, "Model Depth, Input Resolution, and Training Dataset Size as Predictors of Medical Image Classification Performance", Indo American Journal of Multidisciplinary Research and Review, Volume 9, Issue 2, July - December, Page Number 201-210, 2025.
**Copy Right:** © IAJMRR Publication, 2025 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
**DOI:** https://doi.org/10.5281/zenodo.24039053

**Abstract:**

This analysis examines whether model expansion consistently improves medical image classification, informing international debates on benchmark evaluation. A secondary analysis of the Med MNIST v2 benchmark published in 2023 uses 40 configuration observations across ten datasets, matched architecture and input comparisons, descriptive correlations and exploratory exclusions. Increasing architecture depth yields an average AUC decrease of 0.00285, with improvements and declines each occurring in four datasets. Enlarging inputs increases average AUC by 0.00335 and improves performance in eight datasets. Because enlargement resizes existing images, this comparison concerns input representation rather than additional acquired information. Logarithmic training size correlates positively with mean dataset AUC at 0.491, but the association becomes negligible after excluding the smallest dataset. That exclusion also reverses the average depth comparison, while resizing retains a positive direction across both sensitivity samples. These findings qualify expectations that expansion necessarily improves discrimination by separating matched configuration differences from dataset composition. The contribution supports transparent benchmark aggregation and diagnostic task comparisons before configuration preferences guide model development.

**Key Words:** Medical Image Classification; Benchmark Evaluation; Model Depth; Input Resizing; Sensitivity Analysis

## 1. Introduction:

Selecting a medical image classifier requires evidence about which design choices improve discrimination and whether those improvements persist across diagnostic tasks. Greater model depth, larger image inputs and more training images each offer plausible advantages, but they represent different changes to the learning problem. Treating them as interchangeable indicators of model quality can obscure whether performance differences arise within a diagnostic task or reflect differences between datasets. This distinction matters for researchers choosing benchmark configurations and for evaluation teams interpreting claims of algorithmic improvement. Representation learning provides the conceptual basis for examining these choices. Successive network layers transform image information into increasingly abstract features, allowing models to represent patterns relevant to classification. LeCun et al., 2015, explain how this hierarchical organisation supports learning from complex data. Additional depth can expand representational capacity, but capacity alone does not establish better discrimination on an unseen test set. Medical imaging applications differ in their visual characteristics, task definitions and available training information. The synthesis by Litjens et al., 2017, demonstrates the breadth of these applications and the importance of evaluating learning approaches within their particular diagnostic settings. Architecture comparisons therefore require attention to the task over which performance is measured.

Input dimensions introduce a related but distinct issue. Enlarging an image changes the spatial representation processed by a network, whereas acquiring an image at greater native resolution may provide additional diagnostic information. These operations should not receive the same interpretation. The Med MNIST v2 benchmark published by Yang et al., 2023, evaluates residual network architectures with 18- and 50-layers using inputs of 28 by 28 and 224 by 224 pixels. The larger inputs in these comparisons were resized from the smaller images. Consequently, their performance differences provide evidence about sensitivity to input representation rather than the value of additional acquired anatomical detail. The benchmark supplies matched configurations through which architecture and resizing comparisons can be distinguished. Training dataset size operates at another analytical level. Additional informative examples can support parameter estimation, but image count does not directly measure diagnostic diversity or task difficulty. Comparing different training counts across unrelated classification tasks therefore differs from increasing the training sample for a fixed task. Varoquaux and Cheplygina, 2022, identify dataset characteristics and evaluation practices as central challenges in medical imaging research. Their analysis motivates a distinction between a descriptive association across datasets and

evidence that additional training images independently improve performance. That distinction is especially important when a benchmark contains several configurations evaluated on the same training split. Repeated configurations increase the number of performance records without increasing the number of distinct training samples available for a comparison between datasets.

Outcome interpretation also requires a defined boundary. Test area under the receiver operating characteristic curve, abbreviated as AUC, measures discrimination within the specified evaluation sample. It does not establish that a configuration will retain its performance in another institution or improve clinical decisions. Zech et al., 2018, demonstrated variation in pneumonia detection performance across hospital systems, showing why predictive success depends on the evaluation setting. Kelly et al., 2019, similarly distinguish algorithmic performance from the broader requirements for clinical impact. The present analysis therefore uses published test AUC as a benchmark discrimination measure and confines its conclusions to the evaluated configurations and tasks. The closest empirical foundation is the Med MNIST v2 benchmark itself, which reports configuration specific performance across medical image datasets. This analysis extends the interpretation of those published results by separating matched architecture and resizing comparisons from the association between training size and mean dataset performance. Its bounded contribution is to examine whether aggregate configuration preferences persist across diagnostic tasks and whether descriptive conclusions depend on benchmark composition. This is a secondary analysis of existing measurements rather than a new image collection, model architecture or training experiment.

The empirical sample comprises 40 configuration observations across ten binary and multiclass medical image datasets from the benchmark published in 2023. Each dataset contributes four configurations, enabling comparisons of architecture at a fixed input dimension and comparisons of resizing within a fixed architecture. Training size is examined separately through its logarithmic transformation and mean AUC across configurations for each dataset. Descriptive correlations and exploratory sample exclusions assess the influence of related organ datasets and the smallest training dataset. This design makes the source of comparison explicit without attributing differences between diagnostic tasks to a single predictor. This study aims to determine how architecture depth, input resizing and training dataset size relate to published medical image classification performance. The analysis asks how AUC differs between architectures with 18 and 50 layers, how it changes when inputs are enlarged from 28 to 224 pixels per side, and how mean dataset AUC varies with logarithmic training size. The completed comparisons show a predominantly positive resizing pattern, an inconsistent depth comparison and a training size association sensitive to the smallest dataset. These findings qualify a general expectation that larger configurations or datasets necessarily produce better benchmark performance. Their practical contribution is to support matched diagnostic task comparisons and transparent sensitivity reporting before an overall configuration preference is presented.

## 2. Theory, Conceptual Framework and Research Questions Development:

Medical image classification depends on how model capacity, image representation and training information interact. Representation learning provides the theoretical foundation: successive processing layers transform image features into representations that support discrimination between classes. LeCun et al., 2015, explain how these hierarchies enable models to learn increasingly abstract features. This analysis examines three observed dimensions of that process: model depth, input resolution and training dataset size. Classification performance is measured by test area under the receiver operating characteristic curve, abbreviated as AUC. The framework concerns associations with benchmark discrimination rather than clinical effectiveness. Model depth represents the number of layers in the evaluated residual network architecture. Additional layers can support more complex feature transformations, potentially improving discrimination when diagnostic patterns require combinations of spatial and morphological features. However, additional capacity does not ensure better generalization when the available training information cannot support its estimation. Litjens et al., 2017, document the diversity of medical imaging applications and the importance of matching learning approaches to their data and tasks. Consequently, depth provides a theoretical opportunity for improvement rather than an established directional relationship across the included datasets. The available observations compare architectures with 18 and 50 layers at both input resolutions within each dataset. These matched comparisons hold dataset identity and input resolution constant. They therefore support an evaluation of architecture differences without treating variation between diagnostic tasks as evidence of a depth advantage. Because the architectures differ in more than their layer counts, any observed contrast is interpreted as an architecture comparison indexed by depth.

RQ1: How does test AUC differ between residual network architectures with 18 and 50 layers within the same dataset and input resolution?

Input resolution concerns the dimensions of the image presented to the network. Changing these dimensions alters the spatial representation processed by convolutional operations and may change how existing image patterns interact with the architecture. The relevant distinction is between adding diagnostic information and changing the representation of information already available. Yang et al., 2023, specify that the benchmark inputs of 224 by 224 pixels were resized from images of 28 by 28 pixels.

Enlargement therefore cannot be interpreted as recovering anatomical detail lost during initial image reduction. This analysis compares the two input dimensions within each dataset and architecture. A performance difference would indicate sensitivity to input representation under the published training conditions. Improvement, deterioration and negligible change are all admissible outcomes because resizing can alter feature processing without increasing the underlying information content. The research question therefore avoids assuming that larger inputs must produce higher discrimination.

RQ2: How does test AUC differ between inputs of 28 by 28 pixels and inputs resized to 224 by 224 pixels within the same dataset and architecture?

Training dataset size represents the number of images available for parameter estimation. More training examples can expose a model to a broader range of relevant patterns and reduce dependence on individual observations. However, image count does not directly measure the diversity, independence or diagnostic informativeness of those examples. Shen et al., 2017, describe medical image learning as the estimation of hierarchical representations from data, establishing why the information available during training constrains what a model can learn. The theoretical expectation of improved generalisation with additional informative examples does not establish that larger datasets must outperform smaller datasets addressing different classification tasks. Training counts remain constant across the four configurations evaluated for each dataset. The available comparison therefore concerns differences between datasets, using the logarithm of training image count and the mean AUC across configurations. This relationship is exploratory because dataset size varies alongside modality, class structure and task difficulty. Varoquaux and Cheplygina, 2022, show why dataset characteristics and evaluation choices can undermine broad conclusions drawn from medical imaging benchmarks. Accordingly, this analysis interprets training size as an observed correlate rather than an independently identified cause of performance.

RQ3: What association exists between the logarithm of training dataset size and mean test AUC across the included datasets?

The conceptual framework links three directly observed predictors to one observed outcome. Depth and input resolution support comparisons within datasets, while training size supports an exploratory comparison between datasets. These links represent analytical associations, not latent constructs or a composite configuration score. The 40 configuration records describe repeated evaluations across ten datasets rather than 40 independent datasets. This structure permits the three research questions to be addressed while retaining the distinction between architecture comparisons, resizing comparisons and differences in training sample availability.

## 3. Data:

This analysis traces published benchmark results to their source tables, documents sample selection and defines the observed predictors and outcome. The resulting dataset supports comparisons of model architecture and input resolution, together with descriptive examination of training sample availability.

### 3.1 Data Sources, Population, and Sample Construction

The numerical source is the published Med MNIST v2 benchmark provided by Yang and colleagues. Table 2 supplies dataset characteristics and training split counts, while Table 3 supplies test AUC for four residual network configurations. The source was published in 2023, within the evidence window from 2010 to 2024. The extraction timestamp is recorded separately in the accompanying workbook. The target population comprises medical image classification benchmarks evaluated using comparable residual network configurations. The accessible sampling frame contains the 12 two-dimensional Med MNIST v2 datasets and their four published configurations: architectures with 18 or 50 layers, each evaluated with input dimensions of 28 by 28 or 224 by 224 pixels. Selection retains binary and multiclass classification tasks. Chest MNIST is excluded because it uses multilabel classification, while Retina MNIST is excluded because its task is ordinal regression. These rules are documented in the supplied workbook; their establishment before inspection of the published results is not independently documented.

The unit of observation is a dataset and configuration combination. The analytical sample contains 40 observations across ten datasets, with four configurations per dataset. It is a complete configuration cross section rather than an annual panel. The empirical setting covers biomedical image classification across pathology, microscopy, radiography, optical coherence tomography, ultrasound and computed tomography. Dataset identity represents a diagnostic task, not a country. Training counts measure images assigned to a training split, while AUC measures published test discrimination.

Table 1: Sample construction

| Stage | Dataset Units | Configuration Observations | Reason or Rule |
|---|---|---|---|
| Initial accessible frame | 12 | 48 | Published two-dimensional benchmark datasets with four residual network configurations |
| Chest MNIST exclusion | 1 removed | 4 removed | Multilabel classification |
| Retina MNIST | 1 removed | 4 removed | Ordinal regression |

| Stage | Dataset Units | Configuration Observations | Reason or Rule |
|---|---|---|---|
| exclusion | | | |
| Eligible sample | 10 | 40 | Binary and multiclass classification |
| Integrated and checked sample | 10 | 40 | Dataset identifiers matched to source characteristics and performance values |
| Retained descriptive sample | 10 | 40 | Complete measured predictors and test AUC |

The source documentation specifies Creative Commons Attribution 4.0 for the Med MNIST collection, except Derma MNIST image data, which carry an Attribution Non Commercial restriction. This analysis extracts published benchmark statistics rather than individual medical images.

**3.2 Data Integration, Cleaning, and Quality Assurance:**

Integration uses dataset name to connect training split information with performance results. Each dataset contributes one training count and four configuration specific AUC values. The analytical key combines dataset name, architecture depth and input dimension. Verification identified 40 unique keys, four observations for every retained dataset and no unmatched dataset identifiers. Repeated training counts reflect the shared training split across configurations and are not duplicate observations. Source priority follows the published article. The workbook records differences between published organ dataset counts and counts subsequently displayed on the project website. The analytical data retain the published counts accompanying the benchmark results: 34,581 training images for Organ AMNIST, 13,000 for Organ CMNIST and 13,940 for Organ SMNIST. This rule prevents numerical integration across different documented versions.

All 40 analytical training counts and AUC values match their corresponding entries in the workbook's source table. Recalculation also confirms the logarithmic training size transformation. These checks establish internal consistency between the analytical and preserved source sheets; they do not constitute independent replication of model training. No required analytical field is missing. Each retained dataset contributes complete depth, input dimension, training count and AUC measurements. No imputation, interpolation, substitution with zero or removal of extreme observations was applied. Training counts are positive, architecture and input dimensions match the documented configurations, and all AUC values fall within the admissible interval from zero to one. The data retain valid performance extremes because differences between diagnostic tasks are substantively relevant. No trimming or winsorisation was implemented. Dataset heterogeneity nevertheless limits comparisons across tasks. Varoquaux and Cheplygina, 2022, explain why medical imaging benchmark differences require interpretation in relation to dataset characteristics and evaluation design.

**3.3 Variable Measurement, Construction, and Validation:**

The analytical framework uses three observed predictors and one observed outcome. Architecture depth and input dimension are configuration attributes, while training size is a dataset attribute. No latent construct or composite configuration score is estimated. Table 2 defines the measurements and their coverage.

Table 2: Variable dictionary

| Variable or Indicator | Analytical Role | Definition and Construction | Source and Identifier | Unit and Frequency | Transformation and Direction | Missing Value Rule | Coverage and Limitation |
|---|---|---|---|---|---|---|---|
| Architecture depth | Predictor | Published residual network depth of 18 or 50 layers | Med MNIST v2, Table 3, method label | Layers per configuration | Untransformed; higher values identify the deeper architecture | Complete; no replacement | 40 observations; architecture differences extend beyond layer count |
| Input dimension | Predictor | Square input side length of 28 or 224 pixels | Med MNIST v2, Table 3 and benchmark methods | Pixels per side per configuration | Untransformed; higher values identify enlarged inputs | Complete; no replacement | 40 observations; inputs of 224 pixels were resized from 28 pixels |
| Training image count | Predictor | Images assigned to the published training split | Med MNIST v2, Table 2 | Images per dataset | Positive count; higher values indicate more training images | Complete; no replacement | Ten distinct dataset counts repeated across configurations |
| Logarithmic training size | Transformed predictor | Base ten logarithm of training image count | Calculated from the training count | Logarithmic count per dataset | A one unit increase represents a tenfold increase in images | Complete; all inputs positive | 40 records but ten dataset level values |
| Test AUC | Outcome | Published mean area under the receiver operating characteristic curve | Med MNIST v2, Table 3 | AUC per dataset and configuration | Untransformed; higher values indicate better discrimination | Complete; no replacement | Reported to three decimals; trial dispersion is not included |
| Dataset | Grouping | Published dataset | Med MNIST | Diagnostic | Categorical | Complete | Ten datasets; three |

| Variable or Indicator | Analytical Role | Definition and Construction | Source and Identifier | Unit and Frequency | Transformation and Direction | Missing Value Rule | Coverage and Limitation |
|---|---|---|---|---|---|---|---|
| identifier | field | name | v2, Tables 2 and 3 | task | identifier | | organ datasets share a source collection |
| Task and class count | Descriptive fields | Published binary or multiclass task and number of classes | Med MNIST v2, Table 2 | Task category and class count | Retained as source descriptors | Complete | Describe task composition rather than a separate estimated mechanism |
| Publication year | Provenance field | Year of the published benchmark | Article metadata | Benchmark release | Constant at 2023 | Complete | Does not identify image acquisition timing |

The implemented training size transformation is defined in Equation 1.

$$L_d = \log_{10} N_d$$

Here, $L_d$ denotes logarithmic training size for dataset d, and $N_d$ denotes its positive training image count. The transformation changes scale without adding observations or correcting differences in task difficulty. Table 3 reports descriptive statistics calculated from the 40 retained records. Standard deviations use the sample definition. Configuration weighting gives each dataset equal representation because every dataset contributes four observations.

Table 3: Descriptive statistics for measured analytical variables

| Variable | Valid Count | Mean | Standard Deviation | Median | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Architecture depth | 40 | 34.00 | 16.20 | 34.00 | 18 | 50 |
| Input side length | 40 | 126.00 | 99.25 | 126.00 | 28 | 224 |
| Training image count | 40 | 43,868.00 | 53,117.94 | 13,470.00 | 546 | 165,466 |
| Logarithmic training size | 40 | 4.2292 | 0.7092 | 4.1291 | 2.7372 | 5.2187 |
| Test AUC | 40 | 0.9580 | 0.0389 | 0.9670 | 0.857 | 0.998 |

The means and medians of depth and input dimension summarise balanced configuration assignments rather than additional evaluated settings. Training count statistics describe repeated dataset attributes and must not be interpreted as 40 independent measurements of sample availability. AUC measures discrimination rather than calibration or clinical utility. Saito and Rehmsmeier, 2015, demonstrate that ROC based evaluation alone can obscure practically relevant performance differences in imbalanced binary classification. Accordingly, the outcome interpretation remains confined to the published discrimination measure.

**3.4 Analytical Framework, Identification, and Reproducibility:**

The implemented analysis is descriptive and comparative. Architecture comparisons subtract the AUC of the 18-layer configuration from that of the 50-layer configuration within each dataset and input dimension. The two resulting differences are averaged within each dataset. Input comparisons subtract the AUC at 28 pixels from that at 224 pixels within each dataset and architecture, then average the two differences within each dataset. These calculations produce ten dataset summaries for each comparison. Training size is examined alongside the arithmetic mean AUC across the four configurations for each dataset. This produces ten descriptive pairs rather than 40 independent observations. Training count does not vary within datasets, so the comparison cannot isolate its contribution from differences in modality, class structure or diagnostic difficulty. No causal training size effect is identified.

The configuration records contain published mean results rather than individual trial outcomes. Consequently, descriptive dispersion across configurations is not interpreted as uncertainty around model training performance. No regression coefficients, confidence intervals or significance tests are reported. This distinction follows the principle articulated by Wasserstein and Lazar, 2016, that substantive interpretation requires more than a threshold-based declaration of statistical significance. Dependence also extends beyond the four configurations within each dataset because the three organ datasets represent different views derived from a shared source collection. Descriptive comparisons therefore retain dataset identities and avoid treating all records as independent replications. No robustness result is claimed beyond the implemented completeness, uniqueness, source consistency and transformation checks. The accompanying workbook preserves the analytical records, source values, variable definitions and selection documentation. Formula links connect training counts and AUC values to their preserved source entries, while the logarithmic transformation remains inspectable. This organisation supports computational traceability, consistent with Peng, 2011, who identifies reproducible computation as a minimum basis for evaluating numerical research claims. Definitions, configuration assignments and transformations remain consistent across all retained observations. Methods

**4.1 Research Design and Analytical Comparisons:**

This analysis applies a descriptive comparative design to published medical image classification benchmarks. The analytical targets are differences in test discrimination between architectures; differences associated with input resizing and the association between training availability and mean

dataset performance. The unit of observation is a dataset and configuration combination. The retained sample contains 40 observations across ten datasets, with four configurations per dataset. Architecture comparisons use variation within the same dataset and input dimension. Resizing comparisons use variation within the same dataset and architecture. These comparisons preserve the diagnostic task and published data split while changing the configuration under examination. They describe benchmark contrasts rather than independently identified causal effects. In particular, comparing architectures with 18 and 50 layers does not isolate layer count from their other architectural differences.

Training availability varies between datasets and remains constant across configurations within each dataset. Its association with performance is therefore evaluated using ten dataset summaries. This comparison cannot separate sample availability from differences in diagnostic difficulty, modality or class structure. The distinction between configuration comparisons and comparisons across tasks follows the evaluation concerns identified by Varoquaux and Cheplygina, 2022. The sample construction and measurement definitions follow Tables 1 and 2. All retained observations enter the configuration comparisons. Aggregation produces one performance summary per dataset for the training size analysis, preventing repeated configurations from being counted as independent observations of training availability.

**4.2 Mathematical Specification and Aggregation:**

Let $A_{dkr}$ denote published test AUC for dataset d, architecture depth kand input side length r. Depth takes values of 18 and 50, while input side length takes values of 28 and 224 pixels. The mean architecture contrast within each dataset is calculated through Equation 2.

$$D_d = \frac{1}{2} \sum_{r\in\{28,224\}} \left[A_{d,50,r} - A_{d,18,r}\right]$$

Here, $D_d$ is the average AUC difference between the deeper and shallower architectures for dataset d. Positive values favour the architecture with 50 layers, while negative values favour the architecture with 18 layers. Averaging across the two input dimensions gives both evaluated resolutions equal weight. The corresponding resizing contrast is calculated through Equation 3.

$$R_d = \frac{1}{2} \sum_{k\in\{18,50\}} \left[A_{d,k,224} - A_{d,k,28}\right]$$

Here, $R_d$ denotes the average AUC difference between enlarged and smaller inputs for dataset d. Positive values favour enlarged inputs. Because enlargement starts from the smaller images, the contrast measures a change in input representation rather than additional acquired image detail. Dataset performance for the training size comparison is constructed through Equation 4.

$$M_d = \frac{1}{4} \sum_{k\in\{18,50\}} \sum_{r\in\{28,224\}} A_{dkr}$$

Here, $M_d$ is mean test AUC across the four configurations for dataset d. The logarithmic training count, $L_d$, follows Equation 1. Pearson correlation is calculated between $L_d$and $M_d$to describe their linear association. Spearman correlation describes their rank association, assigning average ranks to tied values. Reporting both distinguishes linear covariation from consistency in dataset ordering, following Schober et al., 2018. Aggregate configuration comparisons use the arithmetic mean of the dataset contrasts defined in Equations 2 and 3. Medians and counts of positive, negative and unchanged contrasts describe their distribution. Each dataset receives equal weight. Weighting by image count would instead give larger diagnostic datasets greater influence over the configuration comparison.

**4.3 Processing, Computation and Verification:**

Processing begins with the analytical worksheet and its preserved source values. Dataset names connect training counts to performance measurements. The combination of dataset name, architecture depth and input dimension identifies each analytical record. Verification checks completeness, uniqueness and the presence of all four configurations within every retained dataset. Training counts and AUC values are compared with the corresponding preserved source entries. The logarithmic transformation is recalculated from each positive training count. Additional checks confirm that architecture and input dimensions match the documented configurations and that AUC lies between zero and one. The source priority rule retains the published dataset counts accompanying the benchmark results. Descriptive summaries include valid counts, arithmetic means, sample standard deviations, medians and observed ranges. Sample standard deviations use a denominator equal to the valid countless one. Configuration contrasts are then calculated using Equations 2 and 3, followed by dataset means using Equation 4. Correlations use the resulting dataset summaries rather than the original configuration records.

Computations use Python 3.12.14, NumPy 2.3.5 and SciPy 1.17.0. NumPy supplies arithmetic aggregation and logarithmic transformation. SciPy functions pearsonr and spearmanr supply the reported correlation coefficients. Their probability outputs are not used. All calculations are deterministic and require no random seed, iterative optimisation or convergence criterion. No required analytical value is missing, so no deletion for missingness, imputation or interpolation is applied. Valid extreme observations

remain in the primary sample. Published AUC values are retained at their available precision, with rounding applied only when presenting calculated summaries.

**4.4 Sensitivity Analysis, Interpretation and Replicability:**

Two exploratory exclusions assess sensitivity to benchmark composition. The first removes Organ AMNIST, Organ CMNIST and Organ SMNIST because these tasks share a source collection. The second removes Breast MNIST to examine the influence of the smallest training dataset and its comparatively large adverse architecture contrast. These checks were conducted after inspecting the descriptive results and are not presented as prespecified analyses. For each restricted sample, the analysis recalculates the mean architecture contrast, mean resizing contrast and both training size correlations. The organ exclusion retains seven datasets and 28 configuration observations. The Breast MNIST exclusion retains nine datasets and 36 observations. Every measure within a sensitivity comparison uses the same retained datasets. Changes are assessed through direction and magnitude rather than continued statistical significance.

The analysis reports descriptive coefficients without confidence intervals or significance tests. Published mean AUC values do not supply trial uncertainty, and related datasets limit an interpretation based on independent sampling. Zero contrasts consequently indicate equality at the available reporting precision rather than proof of equivalent model performance. The accompanying workbook preserves source entries, analytical records, variable definitions, selection rules and inspectable construction formulas. Equations 1 through 4 and the stated software functions specify the numerical procedures required to reproduce the reported summaries. This organisation supports the computational traceability advocated by Peng, 2011. The methodological contribution is the explicit separation of matched configuration comparisons from dataset level associations, using established analytical tools rather than a new estimator.

**5. Findings:**

This analysis evaluates architecture depth, input resolution and training dataset size in relation to medical image classification performance. Matched comparisons separate configuration differences within datasets from associations between datasets. Exploratory sensitivity checks examine whether shared organ datasets and the smallest training dataset influence the descriptive conclusions. The analytical sample contains 40 configuration observations across ten medical image datasets, with four configurations per dataset and no missing analytical values. The observations derive from the benchmark published in 2023 rather than repeated annual measurements. Published test AUC averages 0.95798 across configurations, with a standard deviation of 0.03886, a median of 0.96700 and a range from 0.85700 to 0.99800. Each observation represents a published mean performance result, not an individual training trial. Yang et al., 2023, provide the underlying benchmark measurements.

Training availability varies substantially across the ten datasets, from 546 images in Breast MNIST to 165,466 in Tissue MNIST. The median training count is 13,470 images. This variation provides a basis for examining differences in sample availability, but each dataset addresses a distinct classification task. Consequently, comparisons within datasets and comparisons between datasets answer different empirical questions. Table 4 presents dataset specific mean AUC and matched configuration differences. The depth comparison averages the difference between architectures with 50 and 18 layers across the two input dimensions. The input comparison averages the difference between 224 pixel and 28-pixel inputs across the two architectures. Every dataset receives equal weight.

Table 4: Dataset performance and matched configuration comparisons

| Dataset | Training Images | Mean Test AUC | Change Associated With Greater Depth | Change Associated With Enlarged Input |
|---|---|---|---|---|
| Path MNIST | 89,996 | 0.98775 | Increase 0.00350 | Increase 0.00250 |
| Derma MNIST | 7,007 | 0.91550 | Decrease 0.00600 | Increase 0.00100 |
| OCTMNIST | 97,477 | 0.95275 | Increase 0.00450 | Increase 0.01050 |
| Pneumonia MNIST | 4,708 | 0.95250 | Increase 0.00500 | Increase 0.01300 |
| Breast MNIST | 546 | 0.87875 | Decrease 0.03450 | Decrease 0.00050 |
| Blood MNIST | 11,959 | 0.99750 | Decrease 0.00100 | No change |
| Tissue MNIST | 165,466 | 0.93150 | No change | Increase 0.00200 |
| Organ AMNIST | 34,581 | 0.99750 | No change | Increase 0.00100 |
| Organ CMNIST | 13,000 | 0.99275 | Decrease 0.00050 | Increase 0.00150 |
| Organ SMNIST | 13,940 | 0.97325 | Increase 0.00050 | Increase 0.00250 |
| Equal dataset average | 43,868 | 0.95798 | Decrease 0.00285 | Increase 0.00335 |

For RQ1, greater architecture depth does not produce a consistent performance advantage. Four datasets favour the architecture with 50 layers, four favour the architecture with 18 layers and two show no average difference at the available reporting precision. The median depth difference is zero, while the

equal dataset average indicates a decrease of 0.00285 AUC units. The contrast between the zero median and negative mean shows that the aggregate result reflects the magnitude of losses rather than a majority of datasets favouring shallower networks. Breast MNIST accounts for the largest depth related difference, with an average decrease of 0.03450 AUC units. The largest improvement is 0.00500 in Pneumonia MNIST. Removing Breast MNIST in an exploratory sensitivity analysis changes the average depth difference to an increase of 0.00067. Thus, the negative aggregate comparison is sensitive to one dataset. The evidence resolves RQ1 as a heterogeneous architecture comparison rather than a general depth advantage. The representational capacity described by LeCun et al., 2015, therefore remains a theoretical capability whose empirical benefit depends on the evaluated setting.

For RQ2, enlarged inputs show a more consistent descriptive advantage. Eight datasets improve, Blood MNIST shows no change and Breast MNIST records a decrease of 0.00050 AUC units. The average increase is 0.00335, while the median increase is 0.00175. Improvements are largest in Pneumonia MNIST and OCTMNIST, at 0.01300 and 0.01050 respectively. These two datasets contribute most strongly to the aggregate gain. The input comparison has the same aggregate direction within both architectures. Mean AUC rises from 0.95770 to 0.96110 for the architecture with 18 layers and from 0.95490 to 0.95820 for the architecture with 50 layers. The corresponding increases are 0.00340 and 0.00330. This similarity indicates that the average resizing advantage is not confined to one evaluated architecture. Because the larger inputs were resized from the smaller images, the result concerns input representation rather than additional acquired diagnostic detail.

The contrast between depth and resizing is substantively relevant. Greater depth produces balanced counts of improvements and declines, whereas enlarged inputs improve performance in most datasets. Nevertheless, the magnitude of the resizing advantage varies considerably. This pattern supports the task dependent interpretation of medical image learning documented by Litjens et al., 2017, rather than a universal configuration ranking. For RQ3, the Pearson correlation between logarithmic training size and mean dataset AUC is 0.491 across ten datasets. The Spearman rank correlation is smaller, at 0.309. Both describe positive associations, but their different magnitudes indicate that the linear relationship is stronger than the ordering of datasets by size and performance.

Individual comparisons qualify the aggregate association. Tissue MNIST has the largest training count but a mean AUC of 0.93150, below several smaller datasets. Blood MNIST achieves a mean AUC of 0.99750 with 11,959 training images, while Breast MNIST combines the smallest training count with the lowest mean AUC. Training availability therefore does not establish a monotonic performance ordering across diagnostic tasks. The role of training information described by Shen et al., 2017, is consistent with considering sample availability, but the observed comparison does not isolate its contribution from differences between tasks. Table 5 reports exploratory sensitivity analyses conducted after inspecting the descriptive results. Excluding the three organ datasets addresses their common source collection. Excluding Breast MNIST examines the influence of the smallest training dataset and the largest adverse depth comparison. Both checks use the same retained observations for all reported measures within each sample.

Table 5: Exploratory sensitivity of configuration differences and training size associations

| Analytical Sample | Datasets | Configuration Observations | Mean Depth Change | Mean Input Enlargement Change | Pearson Correlation | Spearman Correlation |
|---|---|---|---|---|---|---|
| Complete sample | 10 | 40 | Decrease 0.00285 | Increase 0.00335 | Positive 0.491 | Positive 0.309 |
| Excluding three organ datasets | 7 | 28 | Decrease 0.00407 | Increase 0.00407 | Positive 0.552 | Positive 0.429 |
| Excluding Breast MNIST | 9 | 36 | Increase 0.00067 | Increase 0.00378 | Negative 0.011 | Positive 0.050 |

Excluding the organ datasets preserves the average directions of both configuration comparisons and leaves the training size association positive. Excluding Breast MNIST produces a different result: the depth comparison changes direction, and both training size correlations become close to zero. The resizing comparison remains positive under both exclusions. Accordingly, the resizing pattern is more stable across these sample definitions than either the aggregate depth comparison or the training size association. These sensitivity results identify the scope of the empirical conclusions. RQ1 yields no consistent depth advantage, RQ2 identifies a predominantly positive resizing comparison and RQ3 yields a positive descriptive association that is sensitive to the smallest dataset. Published rounding limits interpretation of small differences, while the available mean results do not establish trial uncertainty or statistical significance. This calibrated interpretation follows the benchmark evaluation concerns raised by Varoquaux and Cheplygina, 2022.

**6. Discussion:**

Greater model complexity does not provide a dependable route to better medical image discrimination. Table 4 shows that the depth comparison is evenly divided between improvements and declines, whereas input enlargement improves AUC in eight of ten datasets. The contribution of this analysis is the separation of these configuration patterns from the association between training availability and performance. The theoretical capacity of deeper networks to learn complex representations, discussed by Schmidhuber, 2015, does not establish that additional depth will improve every diagnostic task. Here, the negative average depth comparison changes direction when Breast MNIST is excluded, as Table 5 demonstrates. This sensitivity changes the interpretation from a general disadvantage of depth to a configuration comparison influenced by one diagnostic dataset. The evidence therefore favours task specific evaluation over architecture rankings based on a pooled mean. Input enlargement provides a more consistent descriptive pattern than additional depth, but its interpretation depends on the information supplied to the network. The larger inputs were resized from the smaller images, so the improvement in Table 4 cannot represent newly acquired anatomical detail. Instead, the comparison indicates that the representation of existing image information matters for benchmark discrimination. Similar average improvements within both architectures further suggest that this pattern is not confined to one evaluated depth. This distinction complements the findings of DeGrave et al., 2021, who demonstrated that predictive performance can depend on image characteristics other than the intended pathological signal. Their evidence concerns shortcut learning, which is not measured here; the relevant connection is that improved discrimination alone does not identify what information a model uses. This analysis consequently distinguishes a reproducible configuration comparison from an explanation of its internal mechanism.

Training availability contributes a different theoretical signal because its variation occurs between diagnostic datasets rather than within matched configurations. The positive correlation using logarithmic training size, constructed through Equation 1, becomes negligible when Breast MNIST is excluded, as Table 5 shows. Larger datasets therefore do not establish a stable ordering of discrimination performance across these tasks. The important distinction is between having more observations and having observations that resolve a particular classification problem. Beam and Kohane, 2018, place medical machine learning within a statistical framework that requires calibrated expectations about what data and algorithms can achieve. The present comparison sharpens that perspective by showing how an apparent training size relationship can depend on the composition of the benchmark collection. Its contribution is an empirical warning against interpreting a cross-dataset correlation as a learning curve for a fixed task. Benchmark composition also shapes which configuration conclusions appear stable. Removing the related organ datasets preserves the average directions of the depth and resizing comparisons, while removing Breast MNIST reverses the depth comparison. Table 5 thus reveals two distinct concerns: shared source datasets affect representation within the benchmark, and an influential diagnostic task can determine the direction of an aggregate result. Neither concern is resolved by increasing the number of configuration records because repeated evaluations do not create additional independent diagnostic settings. Zech et al., 2018, demonstrated variation in pneumonia detection performance across hospital systems, illustrating how evaluation context changes the meaning of apparent predictive success. The present evidence addresses variation across benchmark tasks rather than hospitals, extending the practical lesson that performance summaries require explicit attention to the units over which they are aggregated.

The international relevance lies in improving how medical image models are compared, rather than establishing differences between national healthcare systems. Tables 4 and 5 support evaluating shallower architectures alongside deeper alternatives, assessing resizing within each diagnostic task and reporting sensitivity to benchmark composition. They do not establish computational savings or clinical benefit. Kelly et al., 2019, distinguish algorithmic performance from the wider requirements for clinical impact; this analysis contributes to the evaluation stage by identifying which descriptive conclusions survive changes in the comparison sample. The theoretical refinement is therefore conditional: architecture capacity, input representation and sample availability operate at different analytical levels and should not be treated as interchangeable routes to improved performance. For researchers and evaluation teams, the practical implication is to retain matched task comparisons and disclose influential datasets before presenting an overall configuration preference.

**References:**

1. Beam, A. L., & Kohane, I. S. 2018. Big data and machine learning in health care. JAMA, volume 319, issue 13, pages 1317 to 1318. https://doi.org/10.1001/jama.2017.18391
2. DeGrave, A. J., Janizek, J. D., & Lee, S. I. 2021. AI for radiographic COVID 19 detection selects shortcuts over signal. Nature Machine Intelligence, volume 3, issue 7, pages 610 to 619. https:// doi.org/10.1038/s42256-021-00338-7
3. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, volume 17, issue 1, Article 195. https://doi.org/10.1186/s12916-019-1 426-2

4. LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. Nature, volume 521, issue 7553, pages 436 to 444. https://doi. org/10.1038/nature14539
5. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. Medical Image Analysis, volume 42, pages 60 to 88. https://doi.org/10.1016/j.media.2017.07.005
6. Peng, R. D. 2011. Reproducible research in computational science. Science, volume 334, issue 6060, pages 1226 to 1227. https://doi.org/10.1126/science.1213847
7. 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, volume 10, issue 3, Article e0118432. https://doi.org/10.1371/jou rnal.pone.0118432
8. Schmidhuber, J. 2015. Deep learning in neural networks: An overview. Neural Networks, volume 61, pages 85 to 117. https://doi.org/10.1016/j.neunet.2014.09.003
9. Schober, P., Boer, C., & Schwarte, L. A. 2018. Correlation coefficients: Appropriate use and interpretation. Anesthesia & Analgesia, volume 126, issue 5, pages 1763 to 1768. https://doi.org/10.1213/ANE.0000000000002864
10. Shen, D., Wu, G., & Suk, H. I. 2017. Deep learning in medical image analysis. Annual Review of Biomedical Engineering, volume 19, issue 1, pages 221 to 248. https://doi.org/10.1146/annurev-bioeng-071516-044442
11. Varoquaux, G., & Cheplygina, V. 2022. Machine learning for medical imaging: Methodological failures and recommendations for the future. npj Digital Medicine, volume 5, issue 1, Article 48. https://doi.org/10.1038/s41746-022-00592-y
12. Wasserstein, R. L., & Lazar, N. A. 2016. The ASA's statement on p values: Context, process, and purpose. The American Statistician, volume 70, issue 2, pages 129 to 133. https://doi.org/10.1080/00031305.2016.1154108
13. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. 2023. Med MNIST v2: A large scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, volume 10, issue 1, Article 41. https:// doi.org/10.1038/s41597-022-01721-8
14. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross sectional study. PLOS Medicine, volume 15, issue 11, Article e1002683. https://doi.org/10.1371/journal.pmed.1002683