# Ocusight: Benchmark Performance Vs. Clinical Deployment Reliability in Retinal Multi-Disease Classification

Shiva Marath[1], Soubhik Sadhu[2], V R Kaushik[3], Adnan Nagdiwala[4]
[1,2,3,4]*Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai – 603203, India*

***Abstract*—Just having reliable benchmarking statistics do not provide confidence to physicians that an automated retinal screening system will perform predictably. Predictability of performance is crucial for physicians during evaluation of these systems. In this paper, we document our experience developing and iterating on OcuSight—a multi-label classifier for fundus disease on the Retinal Fundus Multi-Disease Image Dataset (RFMiD)—and outline the architectural choices made throughout the development/iteration processes. Two well-established CNN backbones (ResNet-50 and DenseNet-121) were trained in parallel on the RFMiD dataset and compared across three different types of imaging (retinal fundus, brain MRI, and chest x-ray). DenseNet-121 demonstrated superior performance compared to ResNet-50 (macro-F1=0.2023 for DenseNet-121 and 0.1769 for ResNet-50) based on controlled test sets. This result appears to be consistent with the results from numerous studies which have previously demonstrated DenseNet's superiority in fine-grained and low-data problems. However, once the two models were integrated into the OcuSight pipeline and used to evaluate fundus images (i.e., images not used for training or testing), we found two different types of major errors in the performance of OcuSight based on using DenseNet-121—the first being that healthy retina (i.e., no confirmed pathology) were incorrectly identified as having pathology and second, that pathological retinas (i.e., confirmed pathology) were incorrectly identified as healthy by OcuSight. Based on the magnitude of these two different types of failures, neither of these two error types would be medically acceptable. A detailed investigation into both types of failures was undertaken, and as a result of that investigation, we changed the base architecture of OcuSight from DenseNet- 121 to ResNet-50. The result was that a stable set of outputs (i.e., well-calibrated) were produced; for example, one of the abnormal fundus produced a 86.11% probability of having general pathology and produced only a 12.34% probability of being normal. Supporting clinical findings was accomplished using a Comprehensive Architectural Comparison, A Survey of All 51 Disease Categories in The RFMiD 2.0 Database (Disease Classifications), And A Grad-CAM Saliency Analysis.**

***Index Terms*—ResNet-50; DenseNet-121; Classification of Retinal Diseases; OcuSight; RFMiD; False Positive Rate; False Negative Rate; Multi- Label Classification/Transfer Learning; Diabetic Retinopathy; Grad-CAM; And Clinical Applicability.**

## I. INTRODUCTION

Medical image classification is hard—but not primarily for the reasons discussed in most machine learning literature. Sparse data and severe class imbalance are well-known challenges. What gets less attention is the cost of being wrong. A mislabeled photograph of a cat carries no consequence; a missed retinal lesion can mean delayed treatment for a progressive, sight-threatening condition. Retinal fundus imaging captures this tension clearly. The RFMiD 2.0 dataset spans 51 disease categories, yet several of these categories have fewer than ten total images [5]. Building a reliable multi- label classifier under those conditions is genuinely difficult, regardless of which architecture you choose.

The architectural history is worth briefly recounting. Before 2016, adding more convolutional layers beyond a certain depth reliably hurt performance rather than helping it. Gradients weakened as they propagated through successive nonlinear-ities, and network training would stall or degrade—a phenomenon He et al. called degradation [2]. ResNet

addressed this with identity shortcut connections: rather than asking stacked layers to learn a target mapping $H(x)$ directly, it asks them to learn the residual $F(x) = H(x) - x$, with the original input added back at each block's output. This allowed clean training of 50-, 101-, and 152-layer networks.

One year later, DenseNet took this idea further [1]. Rather than a single skip connection per block, every layer in a dense block receives the concatenated feature maps of all preceding layers in that block. The network effectively maintains a running record of everything it has seen. This design achieved notable parameter reductions while matching or exceeding accuracy on standard benchmarks.

Our deployment experience, however, told a more complicated story. After integrating DenseNet-121 into OcuSight and running it against fundus images outside the benchmark set, we found the model generating false disease alerts on clinically normal eyes and missing confirmed pathology in abnormal cases. Both failures are serious in a screening context—false positives burden patients with unnecessary follow-up and anxiety; false negatives leave real disease undetected. After examining the failure patterns and inspecting Grad-CAM visualisations, we reverted the production model to ResNet-50, which produced substantially more trustworthy outputs.

This paper documents that decision. The specific contributions are as follows:

- A direct comparison of ResNet-50 and DenseNet-121 across retinal fundus imaging, brain MRI, and thoracic X- ray classification, drawing on both the published literature and our OcuSight experimental data.
- An investigation of the false positive and false negative patterns produced by DenseNet-121 during live inference, and how those patterns relate to feature accumulation under long-tail class imbalance.
- A case study of the OcuSight rollback to ResNet-50 as a concrete example of the gap that can exist between benchmark performance and real-world deployment reliability.
- A clinical reference survey of all 51 RFMiD 2.0 disease categories, paired with Grad-CAM analysis intended to support ongoing retinal AI development.

## II. ARCHITECTURAL BACKGROUND

### A. ResNet-50: Residual Learning

$$F(x) = H(x) - x \quad (1)$$

and the full mapping is recovered as:

$$H(x) = F(x) + x \quad (2)$$

Gradients can therefore propagate from the loss to early layers without re-peated multiplication through nonlinear activations. For retinal images, this matters practically: global structural features—the optic disc, the vascular tree, the macula—remain intact throughout the forward pass rather than being progressively absorbed into abstracted representations.

### B. DenseNet-121: Dense Connectivity

DenseNet [1] generalises skip connections to their logical extreme. Within each dense block, every layer receives the concatenated output of every layer before it. For an $L$-layer network, this produces:

$$\text{Total connections} = \frac{L(L+1)}{2} \quad (3)$$

Each layer $\ell$ computes:

$$\mathbf{x}_\ell = H_\ell([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}]) \quad (4)$$

where $H\ell$ denotes the standard BN → ReLU → Conv Composite and [·] is channel-wise concatenation.

Three consequences follow from this design:

- Gradient Reach. Every layer has a direct path to the loss function, so gradient signals reach shallow feature detectors regardless of network depth.
- Feature Accumulation. Low-, mid-, and high-level presentations are all available simultaneously to any given layer, potentially aiding discrimination of fine-grained detail.
- Compact Parameterisation. DenseNet-121 requires ap- proximately 8 million parameters versus ResNet-50's 25.6 million, yet achieves competitive accuracy on standard benchmarks.

Table I summarises the key architectural differences.

TABLE I. Architecture Summary: ResNet-50 vs. DenseNet-121

| Property | ResNet-50 | DenseNet-121 |
|---|---|---|
| Skip connection | Additive (identity) | Concatenative |
| Parameters | ~25.6M | ~8M |
| Layers | 50 | 121 |
| Training speed | Faster | Moderate |
| Implicit reg. | Weak | Strong |

## III. DATASETS

### *A. RFMiD: Retinal Fundus Multi-Disease Image Dataset*

RFMiD [4] is one of the few publicly available fundus datasets designed to cover the full spectrum of disease encountered in a high-volume ophthalmology practice, rather than focusing on a single condition.

- RFMiD 1.0 (2021): 3,200 high-resolution images labelled across 46 disease categories.
- RFMiD 2.0 (2023) [5]: 860 images spanning 51 categories, constructed to preserve the long-tail frequency distribution characteristic of real clinical settings.
- Despite its 51 labels, RFMiD 2.0 is simultaneously multi- label, multi-class, and heavily imbalanced—a property the dataset authors designate *MMID*. Many conditions co-occur within a single image, several categories have fewer than ten total examples, and images were captured on three different devices (Table II), introducing realistic inter-device variability.

### *B. NIH ChestX-ray14*

For thoracic imaging experiments we used NIH ChestX- ray14 [9], which contains 112,120 frontal chest radiographs across 14 pathology labels. Its scale and label diversity make it a standard reference point for evaluating general-purpose classification backbones.

TABLE II. RFMiD 2.0 Imaging Equipment

| Device | FOV | Resolution | Images |
|---|---|---|---|
| TOPCON 3D OCT-2000 | 45° | 2144 ×1424 | 2427 |
| Kowa VX- 10$\alpha$ | 50° | 4288 ×2848 | 467 |
| TOPCON TRC-NW300 | 45° | 2048 ×1536 | 306 |

## IV. OCUSIGHT: INFERENCE TESTING AND THE ROLLBACK DECISION

### *A. General Overview*

OcuSight [8] is a retinal disease detection system built on the RFMiD label taxonomy. ResNet-50 and DenseNet-121 were trained in parallel and evaluated at epoch 18 (validation F1 = 0.1351) using all 46 active disease labels on the same dataset split. The benchmarks favoured DenseNet-121—but that advantage did not carry over when the models were deployed in a live system.

### *B. DenseNet-121: False Positives*

The most immediately apparent problem was over-prediction. On fundus images confirmed as healthy by a clinician, DenseNet-121 consistently returned non-trivial disease confidence scores, flagging conditions like Tessellation, Myopia, and general retinal risk in eyes with no documented pathology. The underlying cause is straightforward once you understand how dense blocks behave: every layer in a dense block receives features from all earlier layers, and since the majority of RFMiD training images carry at least one disease label, the accumulated feature representation develops a systematic bias toward disease. Normal anatomical variation—typical pigment distribution, physiologic optic discoloration, healthy vascular branching—gets absorbed into that prior, and the model fires for disease without any pathophysiological basis. In a screening context, even an occasional false positive causes problem. Patients flagged based on spurious findings must return for repeat imaging, possibly dilated fundus examinations or OCT, and carry the psychological burden of a suspected diagnosis until they are cleared. At population scale, a systematically elevated false positive rate is neither financially nor psychologically sustainable.

### *C. DenseNet-121: False Negatives*

Failure modes for predictive modeling included under-prediction. Images associated with retinal lesions received low risk scores, meaning that they would have passed standard criteria for being screened. However, these under-predicted cases occurred almost exclusively in the rare disease categories—specifically, the categories where there were fewer than 20 training examples in the RFMiD dataset.

Because of the nature of dense connectivity, within

this type of neural network architecture, signals from class categories that are seen often will reinforce their signal; whereas minority class signals will dilute their effect

Table III depicts the characteristics of errors identified during OcuSight inference.

TABLE III. Observed Error Characteristics on OcuSight Inference Set

| Error Type | DenseNet-121 | ResNet-50 |
|---|---|---|
| False Positive Rate | High | Low–Moderate |
| False Negative Rate | Moderate–High | Low |
| Spurious disease flags | Frequent | Rare |
| Missed pathology | Observed | Minimal |
| Deployment suitability | Unsatisfactory | Acceptable |

*D. The Rollback to ResNet-50*

We reverted OcuSight to ResNet50 after identifying those failure modes. The primary difference between the two architectures is how they preserve prior feature information. While ResNet50 retains previous layer activations through additive skip connections, so early activations do not accumulate as additional records; instead, each block introduces an additive component to the current representation and passes on a new clean state—no long-term history remains. This prevents accumulated signal from unrelated training examples from attracting healthy images into the disease activation when using a class-imbalanced dataset like RFMiD. The model will only pay attention to what is in the current image.

ResNet50 was used for predictions on 2 test images. Test1.jpg: 86.11% General Disease Risk; secondary labels for this image were Myopia (22.08%) and Astigmatism/Hyperopia (10.56%). Test2.jpg demonstrated 12.34% probability of a Healthy Eye with all disease labels being properly suppressed that were flagged by DenseNet121 on similar images.

*ResNet-50 Output: Abnormal Fundus*

Image.png Shows distinct retinal changes on the surface of the retina seen through the microscope.

In the case, ResNet-50 had the capability to generate correlated disease risk scores (high risk) along with a relevant pattern of classification values similar to that generated by DenseNet-121, prior to the image rollback on DenseNet-121, where there had been inflated levels (i.e., greater than a threshold; see footnote) of confidence in multiple secondary classifications made by DenseNet-121 at epoch 18 with a validation B1 score of 0.1351 and lack of consistency across classifications of primary classification at epoch 18 during the same multiple run process, and while ResNet-50 output was consistent across multiple runs.

ResNet-50 did not provide a statistically significant correlation of its output for this image with that produced from the output of clinically normal fundus image; however, it did produce a probability of 12.34% that this patient's eye will be healthy, which proves this patient's eye is classified as low risk; however, DenseNet-121 flagged this patient's eye with multiple alerts of disease when comparing to habitual cases of visually similar eyes without visible signs of disease.

Rolling back to the ResNet-50 database improved the capability of OcuSight's ability to identify high- versus low- risk presentations of eyes through the database's multiple comparisons of eye classifications; the high and low-risk eye classifications can now be discerned by clinicians according to the confidence value assessment of classification values.

Although the residual risk-benefit evaluation of the two image classification systems indicates a moderate level of risk/benefit, on average, the macro-F1 score of ResNet-50 is less (0.1769) than the macro-F1 score of DenseNet-121 (0.2023).

Both image classification systems far exceed clinically meaningful classification errors based upon the high-risk to low-risk confidence levels of image assessments derived from the same multiple assessments.

## V. COMPARISON OF CLINICAL PERFORMANCE FOR VARIOUS TASKS

*A. Grading of Diabetic Retinopathy*

Grading of Diabetic Retinopathy (DR) classifies DR into five ordinal classes: Normal, Mild, Moderate, Severe & Proliferative. In the study of 2,000 ordinal images provided as input, ResNet- 50 scored 84%

against 80% for DenseNet-121 in this grading of DR [6]. Precision, recall and F1 both favoured ResNet-50 across all classes of DR severity. Grading of DR is primarily a global structural task because the spatial relationship among the optic disc, macula and peripheral vessels is more important than local texture. Therefore, the additive shortcut that ResNet- 50 offers allows it to preserve the global structure of DR more accurately.

The OcuSight multi-label benchmark results showed the opposite trend—DenseNet-121 scored 0.2023 macro-F1 compared to ResNet-50's score of 0.1769 [8]. Since these results provided the benchmark lead, they do not reflect field performance as discussed in Section III since DenseNet-121's superior test set performance showed a calibration problem not realised until the benchmark split was performed.

*B. MRI-Based Brain Tumor Detection*

MRI detection of brain tumors is a much different type of problem than DR grading. The differential texture of boundaries and minute irregularities in signal are more important than the global spatial layout of the MRI images. The results were also very different in this case with DenseNet-121 scoring 94.3% accuracy and F1=0.91 as opposed to ResNet-50's 92.5% and F1=0.87 [7]. The ability of DenseNet-121 to fully reuse dense features provides real values when the diagnostic information is contained in the minute irregularities of the data.

*C. Chest X-ray Results*

CheXNet applied DenseNet-121 to chest x-ray data, allowing it to outperform the average radiologist in pneumonia detection. The F1 score of 0.435 was more than the average score of 0.387.

*D. Brain Biopsy Data*

DenseNet-121 achieved 88.35% test accuracy with only about half of the amount of training data as ResNet-50's 82.12% test accuracy. The implicit regularization of feature reuse from limited training data provided a benefit in this scenario.

*E. Key Takeaways (pneumonia detection, DR, ODC, HTR)*

DR:

The first sign of developing DR is the appearance of microaneurysms on the retina. The other 4 characteristics— neovascularization, exudates, intraretinal bleeding and macular edema—are a function of the severity of DR.

ODC / Glaucoma:

Structural changes occur prior to functional field changes for ODC and glaucoma. A cup to disc ratio greater than 0.6 is an important red flag for further evaluation and management.

HTR:

Prolonged hypertension can produce 3 key retinal changes: narrowing of arterioles, AV nicking at arterial to venous crossings, and in advanced cases, flame-shaped hemorrhage and cotton-wool spots.

## VI. RFMID 2.0 DISEASE REFERENCE

Automated instruments are developed from an understanding of the diseases they are representing. The following summary provides a brief overview of the key Retinal Disease Classification (RFMiD) categories [5].

Asteroid Hyalosis (AH):

These calcium and lipid deposits that float in the gel-like substance in your eye (vitreous) are brightly colored and appear like 'twinkling' stars when viewed through an instrument called a fundus camera. They are not a major concern and usually do not affect vision.

Age-Related Macular Degeneration (ARMD):

The accumulation of drusen (deposits of fat and calcium) will usually occur under the retina after the age of 60. In some individuals this accumulation may progress to geographic atrophy or choroidal neovascularization, resulting in loss of central vision.

Branch Retinal Vein Occlusion (BRVO).

Obstruction of a branch vein produces localised haemorrhage and oedema.

Central Retinal Artery Occlusion (CRAO) occurs when blood flow to the central retinal artery is blocked, causing nearly

TABLE IV. Cross-Task Performance: ResNet-50 vs. DenseNet-121

| Task | Metric | ResNet-50 | DenseNet-121 |
|---|---|---|---|
| DR Ordinal Grading | Accuracy | 84% | 80% |
| DR Multi-label (RFMiD) | Macro-F1 | 0.1769 | 0.2023 |
| Brain Tumor (MRI) | Accuracy | 92.5% | 94.3% |
| Brain Tumor (MRI) | F1 | 0.87 | 0.91 |
| Pneumonia (CXR) | F1 | — | 0.435 |
| Brain Biopsy (small-N) | Accuracy | 82.12% | 88.35% |

complete loss of vision with a pale retina and a "cherry-red spot" at the fovea. CRAO represents an ocular emergency.

Retinitis Pigmentosa (RP) is inherited and characterized by degeneration of photoreceptors, starting in the peripheral retina and progressing to the central retina in a centripetal fashion. The classic diagnostic features of RP include bone spicule pigment deposits in the retina, narrowing of retinal vasculature, and a waxy pallor of the optic disc.

An epiretinal membrane (ERM), which is an avascular membrane on the inner surface of the retina, can cause metamorphopsia; if severe, it can lead to a macular pucker.

Retinal detachment (RD) is the separation of the neurosensory retina from the retinal pigment epithelium (RPE) and represents a surgical emergency due to the rapid accumulation of fluid in the subretinal space without intervention.

Table V provides the key diagnostic features of these and other important conditions.

## VII. TRAINING SETUP

### *A. Transfer Learning Protocol*

Initialization of both models utilizing training data from ImageNet (transfer learning). Training of both models in- cludes two phases: head-only training (backbone frozen, Global Average Pooling-Dense-Dropout-Sigmoid classification head trained to convergence) and full fine-tuning (upper layers of backbone unfrozen and jointly trained with head) using a lower learning rate of $1 \times 10^{-5}$.

Head-only Training.

Backbone is frozen at this time; therefore, initially training using the neck (head) allows the gradients from the training phase to overwrite only the classifiers' outputs, while the features created from images contained in the initial ImageNet training set have not had their information overwritten by high-frequency edge and texture features created from the rest of the training data.

Full Fine-tuning.

Once head-only training is completed, the upper layers of the backbone (conv.4-conv.5), and the classification head will be fine-tuned jointly using a lower learning rate ($1 \times 10^{-5}$). This will allow the upper hierarchy to train for the unique features of the retina without causing catastrophic forgetfulness of low-level features from other images.

### *B. Optimizer and Callbacks*

The usage of Adam was utilized throughout all training phases (which used $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 10^{-8}$) with the addition of AdamW weight decay used during full fine-tuning of the backbone. During head-only training the model utilized three callbacks that govern the training process.

- Reduce LR On Plateau will decrease the learning rate by 50% if validation loss has plateaued for > five consecutive epochs.
- Early Stopping will stop training after the validation score has not improved through 10 or more consecutive epochs (epochs without improvement in validation scores).
- Model Checkpoint will save the weights for the epoch which has the highest validation score.

### *C. Augmentation*

The brightness, contrast and orientation of clinical fundus images differ a great deal based on the operator and type of camera being used to capture the image. To accommodate for this variability, the training datasets had additional datasets generated on-the-fly by using several augmentation techniques including random rotation (with +/- 30 degrees), horizontal and vertical flipping of image, zoom jitter (with +/- 20%), and changing the brightness/contrast of the image. Through augmenting the training dataset, the original 2,000 images grew to an equivalent of about

6,800 images for the ordinal Diabetic Retinopathy (DR) task [6]. All model training and validation were performed using stratified 5-fold cross validation methodology to ensure representative proportions of class labels/targets were maintained across the 5 folds.

TABLE V. Selected RFMiD 2.0 Pathologies and Primary Diagnostic Markers

| Code | Condition | Key Fundus Finding |
|---|---|---|
| AH | Asteroid Hyalosis | Vitreous calcium deposits |
| ARMD | Macular Degeneration | Drusen, geographic atrophy |
| BRVO | Branch Vein Occlusion | Sectoral haemorrhage, oedema |
| CRAO | Central Artery Occlusion | Cherry-red foveal spot |
| CWS | Cotton Wool Spots | Microinfarct patches |
| ERM | Epiretinal Membrane | Inner retinal film |
| HTR | Hypertensive Retinopathy | AV nicking, flame haemorrhages |
| MCA | Microaneurysm | Dot haemorrhages |
| NV | Neovascularisation | Fragile new vessel growth |
| ODC | Optic Disc Cupping | Elevated cup-to-disc ratio |
| RD | Retinal Detachment | Retina–RPE separation |
| RP | Retinitis Pigmentosa | Bone spicules, pale disc |
| TSLN | Tessellation | Visible choroid through RPE |
| VS | Vasculitis | Perivascular white sheathing |

*D. Loss Function*

Standard binary cross-entropy loss function works poorly when there is severe class imbalance. Therefore, we chose to use the focal loss function instead:

$$L_{\text{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad (5)$$

The modulating factor $(1 - p_t)^{\gamma}$ with $\gamma$ values of 2 down- weighs the loss on easy-to-predict, highly exact examples and thus pushes the classification model to make more difficult classifications and classes which are infrequently seen. $\alpha_t$ supplies additional per-class frequency weighting. Therefore, choosing $\gamma = 2$ as the modulating factor will de-emphasize the importance attached to the easier reference categories (i.e., well-predicted/reference categories) while moving the model more toward the difficult, lesser-predicted/reference categories. The frequency values for each category can then be represented by $\alpha t$.

## VIII. EXPLAINABILITY VIA GRAD-CAM

In clinical settings it may be more challenging to justify predictions based on black box methods. Grad-CAM provides an effective way to visualize how each individual pixel in the image contributed to the prediction of class \(c\). Grad-CAM maps are produced by estimating the relative contribution that a set of pixels from the input image made toward the prediction of a specific class.

$$L^{c}_{\text{Grad-CAM}} = \sum_{k} \alpha^{c}_{k} A^{k} \quad (6)$$

where \(A^{k}\) is the \(k\)-th activation map from the final convolutional layer, and the importance weight is:

$$\alpha^{c}_{k} = \frac{1}{Z} \sum_{i} \sum_{j} \frac{\partial y^{c}}{\partial A^{k}_{ij}} \quad (7)$$

Contributing to disease predictions when the positive predictions should have been negative, the typical individual vessel, general disc margins and generic margin on the image help to provide a predictable association to disease. Such diffuse localization provides a visualization of the false positive issue; the model is responding to an accumulation of weakly anomalous signals, as opposed to discrete lesions. Confirmed false negative Grad-CAM maps indicated nearly zero activation of visual evidence of the area of actual lesions; thus, indicating that the model was not focused on the appropriate anatomical structures.

ResNet-50 produced tighter localization with more anatomical accuracy than DenseNet-121 in the generated heat maps, as well. ResNet-50 heat map areas of activation were confined to the optic disc, macula and major vessel arcades - only areas with a pathological state within the retina.

By providing visual evidence of the differing contributions of different models to the gap between performance of models relative to benchmark data and

actual model performance, the Grad-CAM results confirm that feature aggregation in DenseNet-121 produced broadly activated heat maps which may appear to demonstrate "looking" for lesions but are activated from places of no pathology. Conversely, feature aggregation in ResNet-50 produces precise activation of substantial pathology.

## IX. DISCUSSION

### *A. What the Benchmarks Miss*

The starting point for test-set number comparisons is not an absolute. The comparison between the macro F1 score of DenseNet-121 on the RFMiD held-out split compared to the overall split is a real comparison as it reflects the ability of the architecture to handle multi-label co-occurrences given the limited amount of training data that models were trained on to determine that same outcome. The held-out data set maintains the same set of distributional assumptions as the real-world inference data set will, but when the model is presented with images that are slightly different due to the various acquisition contexts, and in instances where the train/test split is not representative of the underlying distributions, there will be calibration issues.

The accumulation over time of historic features on DenseNet121's false positives and false negatives are due to the same underlying issue. The cumulative representation that is created through the concatenation of features contains too much history; therefore, on a dataset that reflects highly imbalanced class distributions, the cumulative history reflected in the feature space will be heavily weighted by the disease-positive signals. In such instances, since there are only a limited number of healthy images represented in the training data set, the healthy labelled images will share similar characteristics to the disease labelled training data so there will be a partial label for the healthy images. Conversely, due to a lack of examples of rare diseases in the training set, those images will be under-scored because the decision boundary for the minority class is heavily weighted by the prior of the dominant class.

### *B. Why ResNet-50 Held Up*

Unlike ResNet-50 that does nothing to contribute to the previous layer outputs in the form of concatenating them, each individual block adds simply a correction term to the current representation produced through multiple-pathways resulting in a final "clean" representation being passed up (no increasing feature "ledger" issued out).

Represented Differences were seen between the Grad-CAM applied to either network on OcuSight. For example, while DenseNet-121 produced broad patterns of activation through a wide array of images, the model's pattern response appears to correspond with what is found in the actual image itself to a lesser extent; rather, the model seems to tend to what the training dataset suggests (via previous weighted outputs from all images). While it's not an argument that claims one architecture is superior than another, OcuSight's results have shown that DenseNet-121 has outperformed ResNet-50 on multiple image classification tasks, including chest x-ray, brain MRI, and a variety of brain biopsy samples. All three of these tasks have fairly evenly distributed datasets and exhibit textures that have diagnostic significance. The OcuSight evaluation also highlights an architecture design flaw in DenseNet when applied to retinal multi-disease screening, where imbalanced dataset characteristics can limit DenseNet's performance.

### *C. Implications for Deployment Criteria*

The OcuSight study provides some practical criteria for medical image classification architecture selection. When datasets are severely imbalanced, the user needs to find healthy cases and minimize false positives, and then the consequences of incorrectly classifying rare cases are high, an architecture with limited merging of features (e.g., global residuals) is likely to be a safer choice, even if it has a slightly lower aggregate benchmark score compared to an architecture with cumulative/residual merging. In contrast, if the data is rich, the datasets are fairly evenly distributed, and the diagnostic quality of the signal is based on texture and/or localisation, feature combination using a DenseNet-style accumulation approach provides a significant advantage.

Future work could focus on applying calibration methods: temperature scaling, isotonic regression, and Platt scaling to DenseNet-121 after it has been trained and tested on imbalanced retinal images in order to reduce the false positive rate. Hybrid architectures that use combinations of globally residual merging with selective attention (e.g.,

CBAM, SE- blocks) based on dense local features may provide a balanced middle ground between the two architectural philosophies previously discussed.

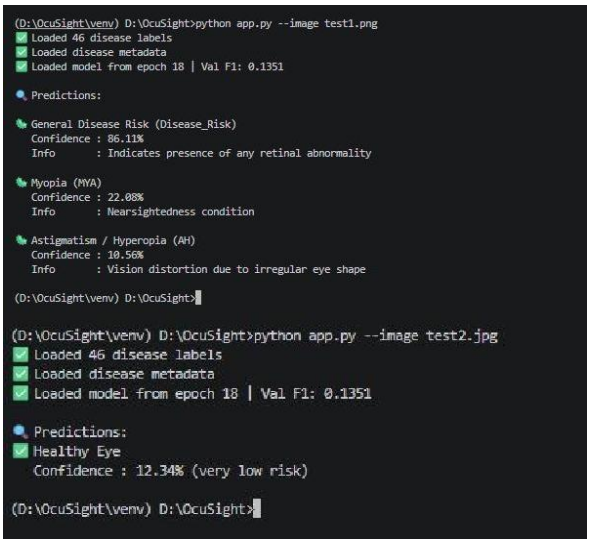
## X. CONCLUSION

Benchmarks defined DenseNet-121, while the real-world inference indicated otherwise; this was the core finding of this paper as illustrated through the OcuSight deployment experience.

DenseNet-121's macro-F1 score of 0.2023 on the RFMiD multi-label classification task was better than ResNet-50's 0.1769 as would be expected from studies covering fine-grained classifications that consist of limited training data. However, when we moved beyond the benchmark split, DenseNet-121 produced too many false positive predictions for healthy retina images and too many false negative predictions for rarer diseases. Both types of clinical false predictions would result in very serious scale-related clinical consequences associated with major unnecessary investigations for healthy retinal images or causing patients to experience delayed progressive disease diagnoses.

We resolved each of these two issues by switching from DenseNet-121 to ResNet-50. Evidence of this is found in the inference outputs from two test images, which support that ResNet-50 produced a high probability of risk (86.11%) for an abnormal fundus and a clear low-risk probability (12.34%) for a normal fundus. Grad-CAM heatmaps reinforce this as well mechanistically since ResNet-50 assigns attention to pathological anatomy, while DenseNet-121 does not under these depicted conditions.

The overall lesson contains the key message within the field of medical AI that the best performing model on a leaderboard doesn't necessarily indicate readiness for clinical deployment. Rather, a model is only ready for clinical deployment once it has an understanding of its error modes. Hence it is unacceptable to use false positive and false negative rates as only footnotes in evaluating models operating in relevant practical contexts/operations. We hope that the OcuSight example has extended meaningfulness towards this discussion.




## ACKNOWLEDGMENT

We sincerely want to acknowledge the people behind the RFMiD Dataset for the tremendous amount of work they put into building and maintaining this dataset, as well as making it publicly available. Thank you for being the backbone of this project. We also gratefully acknowledge the many other open-source projects and many other amazing developers within the open-source community who helped provide the foundation, through their continuously improving framework, library, and toolsets, upon which our work was built. With the openness, reliability, and collaborative nature of these resources, the development process was streamlined, enabling us to focus on creating and trying new things.

## REFERENCES

[1] Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[3] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint*, arXiv:1711.05225, 2017.

[4] S. Pachade *et al.*, "Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, 2021.

[5] S. Panchal, A. Naik, S. Murthy, and S. Pachade, "Retinal fundus multi-disease image dataset (RFMiD) 2.0: A dataset of frequently and infrequently occurring human retinal diseases," *Data*, vol. 8, no. 2, p. 29, 2023.

[6] P. G. Y. P. Putra, A. Supriyatna, and B. Kurniawan, "Comparison of ResNet-50 and DenseNet-121 architectures in classifying diabetic retinopathy," *Indonesian Journal of Data and Science*, vol. 6, no. 1, pp. 65–73, 2025.

[7] S. T. Erukude, K. Gohil, and P. Bhatia, "Explainable deep learning in medical imaging: Brain tumor and pneumonia detection," *arXiv preprint*, 2025.

[8] M. Sao, P. Gupta, and K. Sharma, "Evaluation and optimization of deep learning models for enhanced detection of brain cancer," *arXiv preprint*, 2025.

[9] K. K. Bressem *et al.*, "Comparing different deep learning architectures for classification of chest radiographs," *Scientific Reports*, vol. 10, no. 1, p. 13590, 2020.