PRECISe : Prototype-Reservation for Explainable classification under Imbalanced and Scarce-Data Settings

Vaibhav Ganatra T-VAGANATRA@MICROSOFT.COM
Microsoft Research
India

Drishti Goel T-DRGOEL@MICROSOFT.COM
Microsoft
India

Abstract

Deep learning models used for medical image classification tasks are often constrained by the limited amount of training data along with severe class imbalance. Despite these problems, models should be explainable to enable human trust in the models' decisions to ensure wider adoption in high risk situations. In this paper, we propose PRECISe, an explainable-by-design model meticulously constructed to concurrently address all three challenges. Evaluation on 2 imbalanced medical image datasets reveals that PRECISe outperforms the current state-of-the-art methods on data efficient generalization to minority classes, achieving an accuracy of ∼87% in detecting pneumonia in chest x-rays upon training on < 60 images only. Additionally, a case study is presented to highlight the model's ability to produce easily interpretable predictions, reinforcing its practical utility and reliability for medical imaging tasks.

1. Introduction

In recent years, the integration of deep learning techniques in medical and healthcare applications has exhibited remarkable progress, offering promising avenues for enhanced diagnostic and prognostic capabilities. However, the efficacy of these methods relies heavily on large annotated datasets. Accessing and labelling large quantities of medical images bears humongous costs in terms of the time and medical expertise required. Li et al. (2023) conducted a systematic study of over 300 medical imaging datasets reported between 2013 and 2020 and identified *data scarcity* as the major bottleneck in the adoption of deep learning for medical image analysis. The authors advocate for the development of models that can operate effectively with less data. In addition to scarcity, medical imaging datasets are often characterized by severe class imbalance, with a single/few classes constituting majority of the dataset. For example, Fig. 1 shows the class-wise number (and percentage) of datapoints for RetinaMNIST, a dataset for grading the severity of diabetic retinopathy (DR) in the MedMNIST benchmark Yang et al. (2021, 2023). Out of ∼1100 images, ∼500 images belong to Grade-1 DR, whereas only 66 images belong to Grade-5 DR. A neural network must be able to deal with such class-imbalance effectively in order to have a good performance at the task.

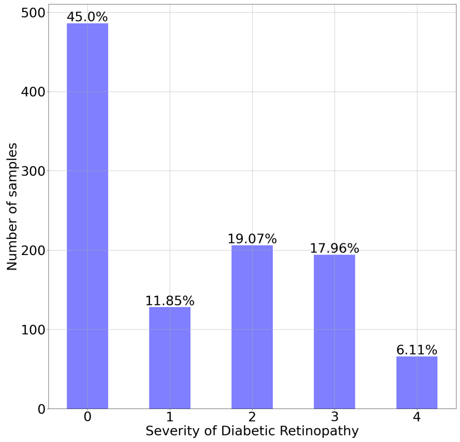


Figure 1: Class-wise data distribution from the RetinaMNIST dataset. Severity-1 Diabetic Retinopathy (DR) makes up 45% of the dataset, whereas only ∼6% (and only 66 images) of the data belongs to Severity-5 DR

In addition to these challenges, for a neural network to be applicable in a medical setting, it is imperative that it can *explain* its predictions. Explainability enhances trust in the model's predictions, which is crucial in a medical context. Current efforts in explainability include *post hoc* methods, which explain the predictions of a trained black-box model by perturbing model-parameters or input/output pairs. Such methods include GradCAMs Selvaraju et al. (2019), Guided BackProp Springenberg et al. (2015), Gradient SHAP Lundberg and Lee (2017) and LIME Ribeiro et al. (2016), among other methods. Jin et al. (2022) evaluate 16 such post-hoc explainability methods on whether they can meet clinical requirements on a multi-modal brain tumour grading task. They conclude that these methods fail to be *faithful* to the model decision process at the feature-level and therefore, cannot be deployed directly in a medical setting. Swamy et al. (2023) also point out the discrepancies among various post-hoc methods when applied to the same input and model, advocating for a shift toward explainable-by-design models for human-centric explainability.

In summary, medical image datasets are small and suffer from heavy class imbalance. Additionally, current post-hoc explainability methods are not reliable for clinical use. Hence, a neural network that is used for medical image classification must be -

- Data-Efficient (It must be able to learn from limited-data)
- Robust to class-imbalance
- Able to provide faithful and human-interpretable explanations

To tackle the aforementioned problems simultaneously, we introduce PRECISe, Prototype-Reservation for Explainable Classification under Imbalanced and Scarce-Data Settings. In summary, we make the following contributions -

- We propose PRECISe, an explainable-by-design neural network that works well with limited and highly imbalanced training data

- We extensively evaluate the performance of our framework on various aspects - overall performance, data-efficiency and robustness to class-imbalance.
- We provide case-studies to demonstrate the explaibability of the proposed method and highlight the ease of human-interpretation of the provided explanations.

Generalizable Insights about Machine Learning in the Context of Healthcare

We propose a framework, PRECISe, for training ML models which generalize well on unseen data with very limited (<60 images) training data. Our solution is robust to the underlying class imbalance present in medical image datasets, and provides human-interpretable explanations. Upon training on 50-60 images, PRECISe achieves $\sim$ 4.5% gains over existing methods. Exaplainability is built into in the proposed framework, which aids in providing consistent and faithful explanations to the human user.

2. Related Work

Data Imbalance and Long-tailed distribution. Several approaches have been proposed to tackle the problem of imbalanced datasets in machine learning. Cao et al. (2019); Li et al. (2021); Mengke Li (2022). Li et al. (2021) attempt to enhance the uniformity of imbalanced feature spaces by distributing features of different classes uniformly on a hypersphere. Mengke Li (2022) propose gaussian clouded logit adjustment via large amplitude perturbation, thereby making tail class samples more active in the embedding space. Cao et al. (2019) propose to replace the cross-entropy loss during training with a label-distribution aware distribution margin loss in an attempt to generalize to tail classes. However, the scale of data on which such models have been trained and evaluated is much larger than typical medical image datasets. Their ability to handle class imbalance on low-data settings needs further evaluation.

Explainable ML in Healthcare. As stated earlier, many of the existing *post-hoc* explainability methods such as GradCams Selvaraju et al. (2019) have been evaluated for their clinical relevance Arun et al. (2021); Ayhan et al. (2022); Van Craenendonck et al. (2020), and have been found to have limited relevance Saporta et al. (2021), with standard methods often highlighting the high-frequency regions in the image Arun et al. (2021). Consequently, alternative explainability methods have been proposed. Boreiko et al. (2022) propose ensembling an adversarial classifier to generate visual explanations for diabetic retinopathy grading. Li et al. (2018) propose a prototype-based model which provides explanations through case based reasoning. However, their utility in scarce and imbalanced data settings has not been validated.

Motivated by this, we propose PRECISe, an explainable-by-design model, which performs well under imbalanced and scarce-data settings.

3. Methods

3.1. Network Architecture

The proposed model consists of 3 components - an auto-encoder, made up of an encoder f and a decoder g, a prototype-metric layer p and a linear classification layer w. The encoder

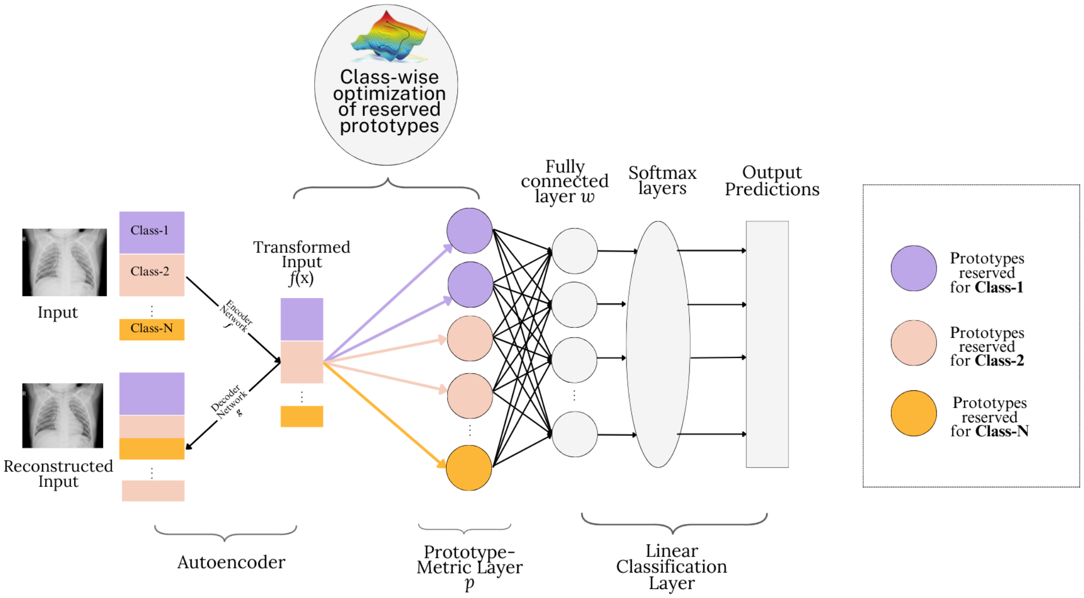

Class-wise optimization of reserved prototypes
Fully connected layer w
Softmax layers
Output Predictions
Input
Class-1
Class-2
Class-N
Transformed Input f(x)
Encoder Network f
Decoder Network g
Reconstructed Input
Autoencoder
Prototype-Metric Layer p
Linear Classification Layer
Prototypes reserved for **Class-1**
Prototypes reserved for **Class-2**
Prototypes reserved for **Class-N**

Figure 2: Overview of Prototype-Reservation

transforms inputs into a compressed lower-dimensional latent space, and the decoder reconstructs the input from the corresponding space. The prototype-metric layer consists of two components - learnt *prototypes* which resemble the representative samples in the training data and a metric layer, which transforms the encoding of an input (under the encoder) into a new *metric* space. It must be noted that the prototypes reside in the latent space of the autoencoder (R^d). An input encoding ($\in R^d$) is transformed into the *metric* space by calculating the Euclidean distance of the encoding from each prototype, i.e. given an input encoding x and m prototypes, the prototype-metric layer transforms x into $\hat{x}$ such that,

$$\hat{x} = p(x) = [\ ||prototype_i - x||_2\ ]\forall i \in [1...m] \tag{1}$$

The output from the prototype-metric layer is an m dimensional vector, where m is the number of prototypes in the layer. The linear classification layer generates a probability distribution over the N output classes from this m-dimensional transformed vector in the *metric* space. It consists of a fully-connected layer and a softmax layer. All the components of the network are trained jointly, as shown in Fig. 2. Given that the prototypes reside in the latent space of the auto-encoder, their visualization becomes straightforward through reconstruction in the input space using the decoder. Explanatory features are inherent in this model, as the distance to each visualizable prototype for any input image can be readily observed. This facilitates a simple evaluation of the decision-making process to ascertain its validity. It is important to emphasize that the architecture of this network closely mirrors the one proposed by Li et al. (2018). However, the novelty of our proposed method lies in the refinement of network optimization, specifically related to prototype synthesis, as elaborated below.

3.2. Network Optimization - Prototype Reservation

Since the network consists of 3 units, the loss function used to optimize the network is also made up of 3 components, the auto-encoder loss, the classification loss and the prototype loss. The auto-encoder loss (AE-Loss) is the (standard) Mean Squared Error Loss, meant to ensure faithful reconstructions of the input from its encodings. Let $D = (x_i, y_i)_{i=1}^{i=n}$, where x_i is the input image and $y_i \in [1...N]$ be the training dataset and $\tilde{x_i}$ is the reconstructed input, corresponding to input x, then the AE-Loss is expressed as -

$$loss_{AE} = \frac{1}{n}\Sigma_{i=1}^{i=n}|| \ \tilde{x_i} - x_i \ ||_2^2; \quad \tilde{x_i} = g(f(x_i) \tag{2}$$

In order to ensure optimal classification performance, the model is optimized with the standard Cross-Entropy loss, i.e. if $\hat{y_i}$ are the probability predictions corresponding to training datapoint (x_i, y_i), the classification loss is expressed as -

Finally, the prototype loss is required to guarantee alignment between the prototypes and training data. This is done by ensuring that all training datapoints are *close* to at least one prototype and that all prototypes are *close* to at least one training datapoint. *Proximity* is considered in the latent space. This idea was proposed by Li et al. (2018). The first component of the prototype loss clusters similar datapoints close to the same prototype and the second component ensures that the prototype is a *summary* of cluster around it. However, adopting this method for imbalanced data leads to all prototypes centered around the majority class, and no prototype corresponding to the minority class. We perform **prototype-reservation** in order to ensure that the network is robust to class-imbalance. Instead of optimizing the prototypes at a dataset-level, we propose optimizing the prototypes at a class-level, i.e. we minimize the distance between prototypes *reserved* for that class and the datapoints corresponding to that class. This ensures that irrespective of the amount of training data, each class is adequately represented in the prototype-metric layer, and therefore robust to class-imbalance. If PR_j represents a set of d prototypes *reserved* for $Class - j$, where d is a hyperparameter, then the prototype-loss is expressed as -

$$L_p = \frac{1}{n}\Sigma_{i=1}^{i=n}\Sigma_{j=1}^{j=N} min_{proto_k \in PR_j} 1[y_i == j].||proto_k - x_i||_2$$
$$+\Sigma_{j=1}^{j=N}\frac{1}{d*N}\Sigma_{proto_k \in PR_j} min_{i \in [1..n]} 1[y_i == j].||proto_k - x_i||_2$$

The first term ensures that all datapoints labelled as j are *close* to at least one of the prototypes ***reserved*** for class-j, whereas the second term ensures that all prototypes ***reserved*** for class-j are close to at least one of the datapoints labelled as j. Running this optimization for all classes present in the dataset ensures prototypes for each class are generated, i.e. $\forall j \in [1..n]$.

The complete loss function is expressed as,

$$loss = loss_{class} + \lambda_1 * loss_{AE} + \lambda_2 * l_p \tag{3}$$

where λ_1 and λ_2 are coefficients used to control the contribution of the auto-encoder and prototype-generation in the model optimization, and are empirically identified. We use $\lambda_1 = 1$ and $\lambda_2 = 0.001$ for all our experiments. A good balance between these 3 terms leads to a network that has visualizable prototypes corresponding to all classes, and therefore, better classification accuracy.

4. Experiments and Results

4.1. Datasets

We evaluate PRECISe on 2 imbalanced medical imaging datasets - the Pneumonia Kermany et al. (2018) and Breast Ultrasound Image (BUSI) Al-Dhabyani et al. (2020) datasets.The pneumonia dataset consists of a total of 5232 Chest X-Ray images from patients (1349 normal, 3883 depicting pneumonia) for training and 624 chest X-rays for testing (234 normal and 390 depicting pneumonia). The dataset consists of AP view pediatric chest x-rays of patients (1-5 yrs) from the Guangzhou Women and Children's Medical Center, Guangzhou. Low-quality / unreadable scans were subsequently removed. The ground truth for the images were obtained by grading by two expert physicians and verified by a third expert. Images are of variable size with the mean size being (1320, 968). The goal is to classify chest X-rays depicting pneumonia from normal images. The BUSI dataset consists of 780 images of breast ultrasound in women, with 487 scans with benign tumours, 210 with malignant tumours and 133 normal scans. The breast ultrasounds were collected from women between 25-75 years of age at the Baheya Hospital for Early Detection Treatment of Women's Cancer, Cairo, Egypt using the LOGIQ E9 ultrasound and LOGIQ E9 Agile ultrasound system. The average size of the images is 500x500 pixels. Images were annotated by radiologists in the same hospital. The small size of these datasets, along with the imbalance in different classes make them ideal for evaluation of PRECISe. While the Pneumonia dataset provides a separate test set, we randomly split 20% of the BUSI dataset and use it as a dedicated test set.

4.2. Baselines and Evaluation

We compare the performance of PRECISe against 3 baselines - a standard ResNet-50 He et al. (2015) model tuned in a fully supervised manner, the original prototypes method proposed by Li et al. (2018) and LDAM Cao et al. (2019), which uses a Label-Distribution Aware Margin Loss to generalize to an imbalanced long-tailed distribution in the training set. As mentioned earlier, we evaluate PRECISe on 3 aspects - data-efficiency, generalization to minority classes and explainability. To estimate the data-efficiency of the proposed method, we take two measures -

- We choose datasets with limited number of images available for training. The scale of these datasets is a fraction of the size of datasets currently used to train neural networks for image recognition such as the ImageNet Russakovsky et al. (2015) or LAION-400M Schuhmann et al. (2021) datasets
- We additionally report performance by training models on subsets of varying sizes (1, 5, 10, 25, 50, 100 % of the entire set). The 1% split is omitted for the BUSI dataset due to its smaller size.

In addition to accuracy, we also report the F1-score averaged over all classes as a performance measure.

To evaluate the generalization of the proposed method to minority classes, we also report the classwise accuracies of all methods. Classwise accuracy is calculated as the proportion of datapoints in each class that the model classifies correctly. We report these for the smallest subset of both datasets (1% subset for Pneumonia, 5% for BUSI) to measure the ability of the model to generalize on minority classes, even with very less data. All results are reported as the mean of 3 independent runs. Different subsets are chosen for the 3 runs when training on <100% of the dataset, however, the subsets are consistent across all methods. Finally, we also present an explainability case-study to demonstrate the utility of visual explanations obtained from the model.

4.3. Implementation

We adopt the ResNet-50 backbone He et al. (2015) for all methods, specifically, the encoder for PRECISe and Prototypes Li et al. (2018) is a ResNet-50 model, with the output projected to a 256-dimensional embedding using a linear layer. The decoder for the proposed method is a fully convolutional network, consisting of upsampling and convolutional operations. Images for both datasets are resized to 224x224 to be used with ResNet50. We do not apply data augmentations for training. We only normalize with the ImageNet mean and standard deviation before processing the image. The number of prototypes reserved for each class was chosen based on experimentation with different number of prototypes reserved for each class (10% data) (details in Appendix). We reserve 2 prototypes for each class in the Pneumonia dataset, and 3 prototypes for each class in the BUSI dataset. Additionally, we use a weighted cross-entropy loss function for the classifier - this helps in preventing the decoder from learning to decode only the majority class. The weights of individual classes are inversely proportional to their frequency of occurrence in the dataset. The model is optimized with the Adam optimizer with a constant learning rate of 1e-3 and a weight decay coefficient of 1e-4. [1]

We empirically found that initializing the architecture with pretrained weights on the ImageNet dataset Russakovsky et al. (2015) yields better accuracies. This has also been observed in prior works Gairola et al. (2021). Possibly, the network's understanding of shape and preliminary structures help in generalization to medical images as well. Hence, all ResNet-50 instances are initialized with pretrained ImageNet weights. Consequently, we call the supervised ResNet-50 baseline as FT (Finetuning) as it is pretrained on the ImageNet dataset and is being finetuned on the current datasets.

4.4. Results and Discussion

Overall Performance : Tables 1 and 2 show the performance of the proposed method and baselines on the Pneumonia and BUSI datasets respectively. The proposed method, PRECISe, outperforms all baselines, achieving the highest accuracy of 92.04% and 88.75% on the Pneumonia and BUSI datasets, respectively.

Data Efficiency. Figures 3 and 4 show the performance of all methods on subsets of varying sizes of the Pneumonia and BUSI datasets, respectively. Again, PRECISe outperforms all methods and achieves highest accuracies. Additionally, whereas other methods suffer

[1] The code implementation for the proposed method is publicly available here

Method	Accuracy	Mean F1 score
FT	89.476 ± 0.421	88.096 ± 0.522
LDAM	87.393 ± 0.659	85.604 ± 0.907
Prototypes	91.293 ± 0.529	90.726 ± 0.526
PRECISe (ours)	**92.041** ± **0.151**	**91.340** ± **0.053**

Table 1: Overall accuracy and Mean F1-scores on the Pneumonia dataset. PRECISe (ours) outperforms all baselines.

Method	Accuracy	Mean F1 score
FT	69.427 ± 0.520	54.143 ± 1.756
LDAM	80.255 ± 0.520	32.718 ± 7.054
Prototypes	87.898 ± 0.520	86.580 ± 0.863
PRECISe (ours)	**88.747** ± **0.601**	**86.939** ± **1.482**

Table 2: Overall accuracy and Mean F1-scores on the BUSI dataset. PRECISe (ours) outperforms all baselines.

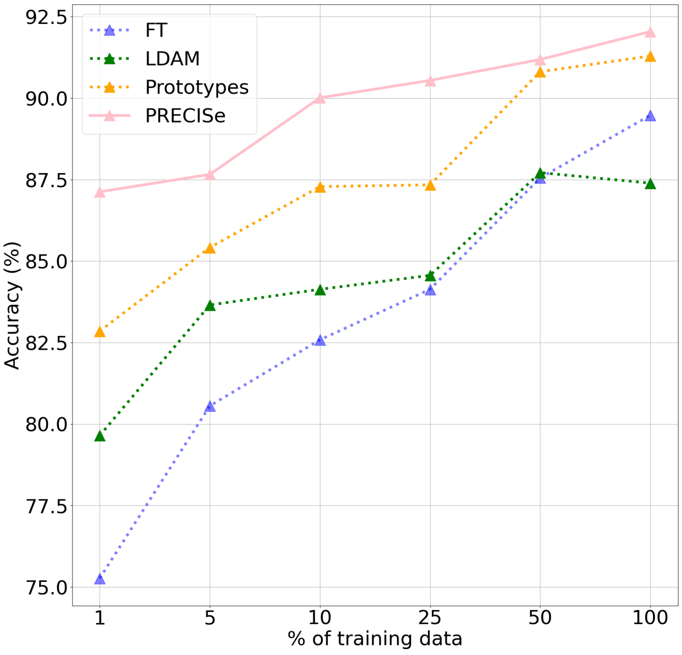


Figure 3: Performance on subsets of varying sizes of the Pneumonia dataset. PRECISe (ours) shows excellent performance retention with reducing training set sizes

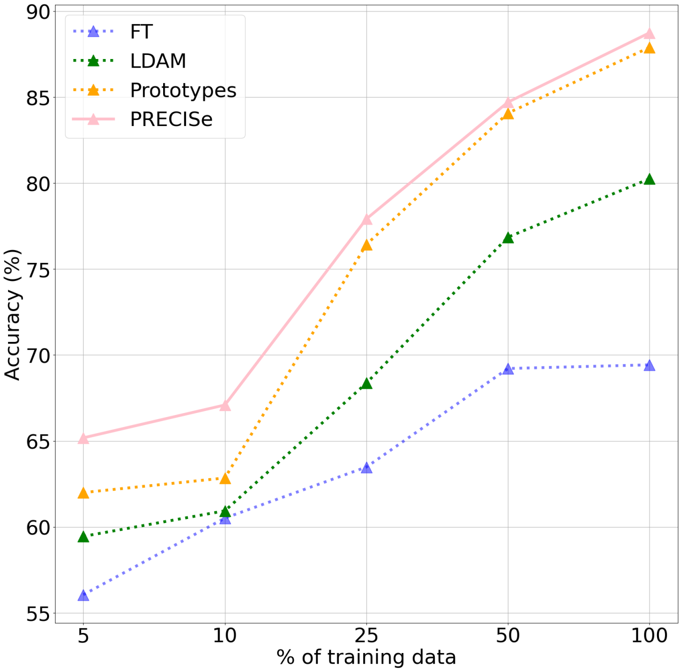


Figure 4: Performance on subsets of varying sizes of the BUSI dataset. PRECISe (ours) shows the best performance at all training set sizes.

from a significant performance drop upon reducing training set sizes, PRECISe shows excellent performance retention. For example, using only 1% of training images (<60) in the Pneumonia dataset, PRECISe achieves an accuracy of 87.13%, as compared to a 92.04% accuracy using the entire Pneumonia dataset. The performance drop is much smaller than that seen in FT (89.48% acc. @ 100% data vs 75.28% acc. @ 1% data) or LDAM (87.39% acc. @ 100% data vs 79.647% acc. @ 1% data). We hope that a model which is able to achieve ∼87% accuracy using <60 labels might go a long way in reducing the time/effort spent by doctors and medical professionals in annotating medical data.

It must be noted that both Prototypes Li et al. (2018) and PRECISe (ours) significantly outperform other methods. We speculate that this is due to the $prototype - metric$ layer. In the typical paradigm of training neural networks, a model is expected to learn representations which compress the information in the input data as well as separate representations of difference classes well. We speculate that the transformation of an input encoding into the *metric* space by finding the Euclidean distance from the prototypes aids the separation of the representations, hence, it is easier for the model to summarize the information about the training distribution in the form of learnt prototypes.

Generalization to Minority Classes - Figures 5 and 6 depict the classwise accuracy of each method on the Pneumonia and BUSI datasets, respectively.

As summarized in Sec 4.1, the “Normal” class is the minority in the Pneumonia dataset, with the majority class having ∼3x data. Similarly, the “Benign” class is the majority class in the BUSI dataset with ∼3x more data than the “Normal” class. In addition to obtaining the best overall accuracies, PRECISe also achieves the best performance on the

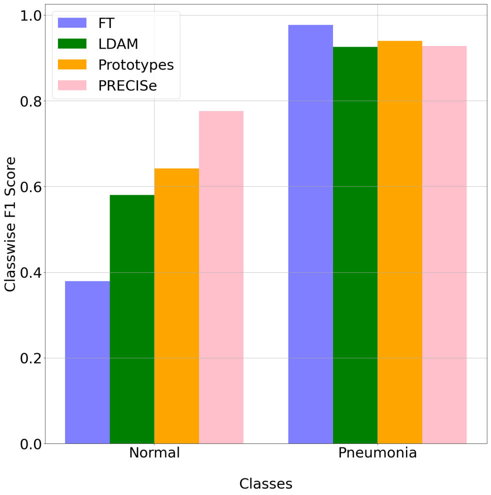


Figure 5: Classwise Accuracy on the Pneumonia dataset

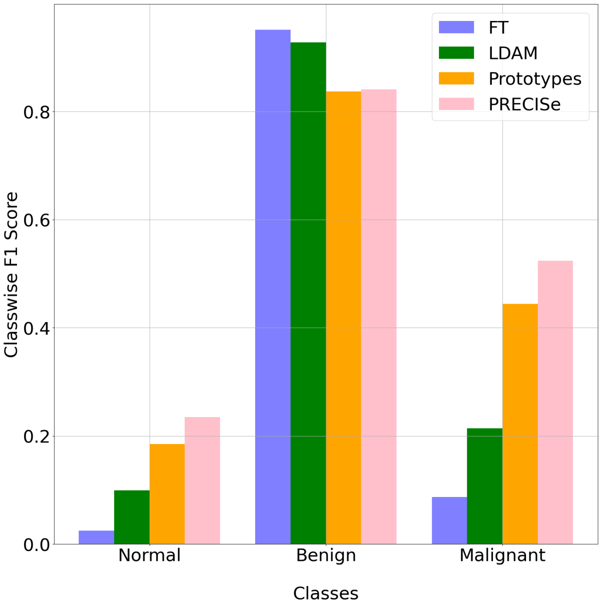


Figure 6: Classwise Accuracy on the BUSI dataset

minority classes, with a comparable performance on the majority class. It must be noted that the results reported for Figures 5 and 6 are for the smallest subset, i.e. 1% subset for the pneumonia set (<60 images) and 5% split for the BUSI dataset (<35 images), while maintaining the class-imbalance ratio. Despite the very small sample size, PRECISe is able to identify 77.6% of all normal chest X-rays of the Pneumonia dataset, as compared to a 37.9% for FT and 58% for LDAM. Similarly, PRECISe is able to identify 52.4% of all Malignant tumours correctly in the BUSI dataset, as opposed to an 8.7% by FT, 21.4% by LDAM and 44.4% by Prototypes. This highlights that fact that improved overall

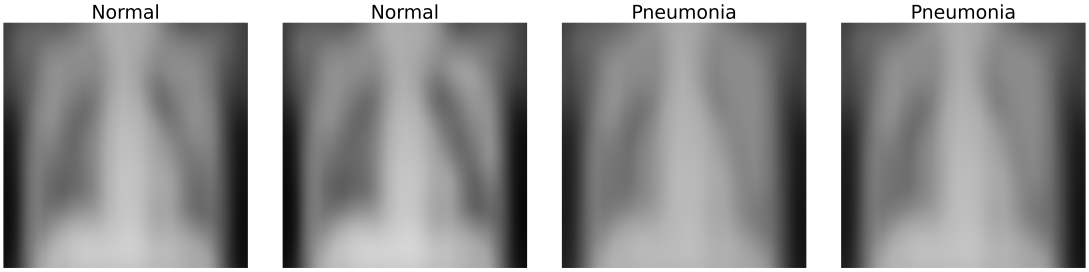


Figure 7: Prototypes generated during the training of PRECISe on the Pneumonia Dataset

performance of the proposed method is because of improved performance on the minority classes, as is desired.

4.5. Explainability Case Study

Fig. 7 shows the prototypes that were learnt when training PRECISe on Pneumonia dataset. Prototypes reserved for normal class appear darker and those of pneumonia appear much paler. This is consistent with the fact that fluid in the lungs manifests as a pale white overlay on the lung region in the chest X-rays. Table. 3 shows the Euclidean distances of randomly chosen normal and pneumonia images (from the test set) on a case-by-case basis from the prototypes. It is distinctly seen that images of either class are much closer to the prototypes of the same class, than those of the other class. Additionally, for each image, the corresponding row is the input to the final linear classification layer. Fully interpretable features being used for classification enables us to examine (and interpret) the basis on the which the model has provided a decision. Moreover, due to the architectural design, the model is optimized to classify based on the "right reasons". This fact is demonstrated in Table. 4, which shows the mean Euclidean distance of all the images in the test set to the prototypes of both classes. Once again, we observe that the datapoints from either class are closer to the prototypes of the corresponding class as compared to the alternative class, which proves the faithfulness of the human interpretable explainations provided by the model.

4.6. Limitations

We acknowledge a few key limitations of the proposed approach. Firstly, the nature of explanations provided by the model are based on whole-image similarity. Unlike existing methods such as GradCams Selvaraju et al. (2019) which highlight the activation regions in the image, PRECISe provides explanations in terms of visually similar images. Hence, interpreting the explanations may still require domain knowledge. Secondly, the proposed setup has not been evaluated by medical professionals. End-to-end evaluation of the approach via user studies may help determine the clinical usability of the system.

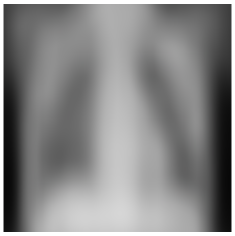
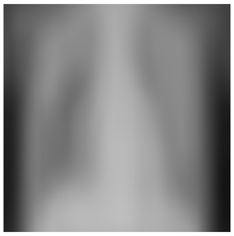
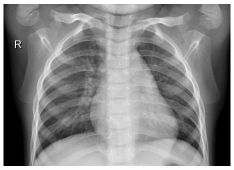
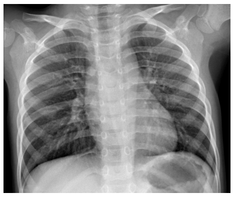
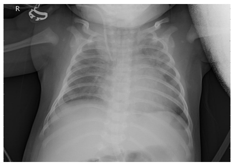
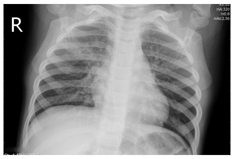
Prototypes → Sample Images ↓	 (Normal)	 (Normal)	 (Pneumonia)	 (Pneumonia)
 (Normal)	**2.283**	**0.519**	7.920	5.424
 (Normal)	**2.363**	**0.712**	7.998	5.499
 (Pneumonia)	7.064	9.501	**1.369**	**3.864**
 (Pneumonia)	6.543	8.962	**1.118**	**3.407**

Table 3: Euclidean distances of images (left) from prototypes (top) - A case-by-case examination of network explainability. Normal images are closer to normal prototypes, whereas images with pneumonia are closer to prototypes for the pneumonia class. For each image, the corresponding row is the input to the final linear layer, thereby using interpretable features for classification.

	Normal Prototypes	Pneumonia Prototypes
Normal Images	**2.536**	5.171
Pneumonia Images	7.932	**3.499**

Table 4: Average distance of all testing data from prototypes reserved for various classes. Data belonging to a particular class is closer to prototypes reserved for that class.

5. Conclusion

In this paper, we proposed PRECISe, an explainable-by-design model which performs well with limited and imbalanced data. We extensively evaluate the model, in various aspects, such as overall-performance, performance on minority classes, data-efficiency as well as the explainations provided by the model. We hope that the proposed method be a first step in the development of holistic-models with a deployment-first mindset, i.e. models which simultaneously tackle multiple problems associated with automated medical image analysis. Future directions of the proposed method include a more thorough examination of the explanation provided by the model, and examining the utility of the learnt prototypes for synthetic data generation.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2019.104863. URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, and Jayashree Kalpathy-Cramer. Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging, 2021.

Murat Seçkin Ayhan, Louis Benedikt Kümmerle, Laura Kühlewein, Werner Inhoffen, Gulnar Aliyeva, Focke Ziemssen, and Philipp Berens. Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. *Medical Image Analysis*, 77: 102364, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102364. URL https://www.sciencedirect.com/science/article/pii/S1361841522000172.

Valentyn Boreiko, Indu Ilanchezian, Murat Seçkin Ayhan, Sarah Müller, Lisa M. Koch, Hanna Faber, Philipp Berens, and Matthias Hein. Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 539–549, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16434-7.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss, 2019.

Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Evaluating explainable ai on a multimodal medical imaging task: Can existing algorithms fulfill clinical requirements? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11945–11953, Jun. 2022. doi: 10.1609/aaai.v36i11.21452. URL https://ojs.aaai.org/index.php/AAAI/article/view/21452.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2018.02.010. URL https://www.sciencedirect.com/science/article/pii/S0092867418301545.

Johann Li, Guangming Zhu, Cong Hua, Mingtao Feng, Basheer Bennamoun, Ping Li, Xiaoyuan Lu, Juan Song, Peiyi Shen, Xu Xu, Lin Mei, Liang Zhang, Syed Afaq Ali Shah, and Mohammed Bennamoun. A systematic collection of medical image datasets for deep learning. *ACM Comput. Surv.*, 56(5), nov 2023. ISSN 0360-0300. doi: 10.1145/3615862. URL https://doi.org/10.1145/3615862.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. *arXiv preprint arXiv:2111.13998*, 2021.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Yang Lu Mengke Li, Yiu-ming Cheung. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, pages 6929–6938, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*, 2021. doi: 10.1101/2021.02.28.21252634. URL https://www.medrxiv.org/content/early/2021/03/02/2021.02.28.21252634.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007%2Fs11263-019-01228-7.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

Vinitra Swamy, Jibril Frej, and Tanja Käser. The future of human-centric explainable artificial intelligence (xai) is not post-hoc explanations, 2023.

Toon Van Craenendonck, Bart Elen, Nele Gerrits, and Patrick De Boever. Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection. *Transl. Vis. Sci. Technol.*, 9(2):64, December 2020.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Appendix

Determining the number of prototypes reserved for each class

As mentioned in Sec 4.3, the number of prototypes reserved for each class was chosen based on experimentation. We tried reserving 1-5 prototypes per class for both the Pneumonia and BUSI datasets using 10% of the training data for experimentation. We found that reserving 2 prototypes per class for the Pneumonia dataset and 3 prototypes per class for the BUSI dataset to yield the best performance. Table 5 shows the performance of PRECISe upon reserving varying number of prototypes. Interestingly, we find that varying the number of prototypes reserved per class does not cause a significant difference in the performance on the Pneumonia dataset, but leads to considerable difference in performance on the BUSI dataset.

#-prototypes	Pneumonia	BUSI
1	88.30	60.51
2	**90.01**	62.42
3	89.90	**67.09**
4	89.26	61.14
5	89.58	62.42

Table 5: Performance of PRECISe (ours) upon reserving different number of prototypes per class.