Hard Sample Mining to Identify Challenging Samples for Breast Cancer Tumor Classification

Mariana Aya S. Uchida[1], **Afonso Matheus S. Lima**[1],
Elaine P. M. de Sousa[1], **Agma J. M. Traina**[1]

[1]Institute of Mathematics and Computer Science (ICMC)
University of São Paulo (USP) – São Carlos, Brazil

{mariaya25, afonso.matheus}@usp.br, {parros, agma}@icmc.usp.br

Abstract. *Deep learning has greatly advanced image classification in medical imaging, aiding the early detection of breast tumors. However, training deep learning models still presents challenges related to model efficiency and biases in data distribution. Breast tumor images highlight these challenges due to their complex tissue structures and overlapping cells. Hard Sample Mining (HSM) has emerged as a promising approach, focusing on selecting representative samples by ranking them based on loss or similarity. This work presents an experimental analysis of the effects of HSM on a binary breast tumor classification task using the BreakHis dataset. The results shows the method's potential for sample ranking in training set pruning.*

1. Introduction

The use of deep learning algorithms has led to significant advancements in computer vision tasks, including image classification, object detection, and image segmentation. However, to enhance the performance of these algorithms, large amounts of data are required. Additionally, it's important to note that not all samples pose the same level of challenge for neural networks to learn from [Xue et al. 2019, Kumar et al. 2024]. To highlight the significance of sample complexity in model training, Hard Sample Mining (HSM) methods have emerged as a promising strategy to identify samples that present greater challenges during neural network training, ultimately improving model performance [Liu et al. 2025a]. In this context, identifying hard samples in medical imaging scenarios is even more crucial, given the need for extensive annotation expertise and the challenge of collecting large, well-annotated datasets for deep learning models to train effectively [Xue et al. 2019, Hesamian et al. 2019]. Pathological images have become an essential resource in clinical practice for the accurate diagnosis of breast cancer, a serious illness ranking fifth globally among causes of death in women [Spanhol et al. 2016, Chen et al. 2025, Chengxiao et al. 2023]. Early detection and accurate diagnosis are crucial for effective treatment and better outcomes in the fight against the illness. However, there are several challenges associated with diagnosing breast cancer accurately. The tissue structure in pathological images is often complex and blurred, making analysis difficult. Additionally, noise factors such as bubbles and overlapping cells can be introduced during specimen preparation and data scanning, further

Acknowledgments. This study was financed in part by the São Paulo Research Foundation (FAPESP – grants 2023/18026-8, 2024/13328-9, 2025/20986-5, and 2026/12389-0.), the National Research Council (CNPq), and the Coordination for Higher Education Personnel Improvement (CAPES – Finance Codes 001).

complicating the classification process [Chengxiao et al. 2023, Chen et al. 2025]. Due to these challenges, a method to identify complex histology samples is desirable.

This study evaluates HSM techniques for identifying hard samples in breast cancer tumor classification tasks. We elaborated the following research questions: **RQ1:** Can HSM techniques for pre-selecting training data achieve comparable performance to deep learning models trained without sample pre-selection in the context of medical imaging? **RQ2:** What are the main characteristics that differentiate hard samples from easy samples in breast cancer tumor classification? Our analysis was conducted on the BreakHis dataset and primarily contributes by providing an experimental methodology for identifying challenging samples in breast cancer histology images and evaluating sample pre-selection for binary classification tasks in medical imaging.

2. Background and Related Works

HSM can be described as a technique that forces models to concentrate more in hard samples within the datasets, rather than treating all the samples equally [Liu et al. 2025a]. The definition of challenging samples can vary depending on the specific task and the criteria used to identify them. Hard samples can be instances in the training set that are particularly difficult for the model to accurately classify. These samples have a high probability of being misclassified, especially when a hyperparameter sets the threshold between easy and hard instances. HSM can be divided into **local** and **global** mining based on the scope of the data samples. Local mining involves training, ranking, and selecting samples within a minibatch to retrain models. On the other hand, global HSM methods consider samples across multiple batches or at the class level to identify hard samples [Liu et al. 2025a]. To address the significance of sample complexity during model training with local HSM, [Liu et al. 2025b] introduced γ-Razor, a two-stage framework that selects representative samples from large-scale datasets, achieving competitive efficiency and converging 5.2x faster than counterparts trained on the original datasets. In a medical imaging context, to address the difficulties of training deep neural networks with noisy-labeled images in a skin lesion classification task, [Xue et al. 2019] defined hard samples as those with ambiguous appearance between benign and malignant cases. Using an online uncertainty-based sample mining method, the authors achieved promising accuracy. In a gland segmentation task, [Li et al. 2022] proposed a novel Online Easy Example Mining method to focus on credible signals, achieving an increased segmentation performance of 4.4% at mean Intersection over Union (mIoU). Based on the existing literature, this study introduces an experimental analysis of local HSM to address hard and easy samples in the context of binary classification of breast cancer histology images.

3. Methodology

Proposed approach. We evaluate local HSM to pre-select samples for a binary classification task. Figure 1 represents the methodology used in this work, with letters *A* to *E* indicating its respective step subsequently.

Dataset. The experiments were conducted using the Breast Cancer Histopathological Image Classification (BreakHis) dataset [Spanhol et al. 2016], which includes 9,109

Regarding generative AI usage. Generative AI tools were used solely for English language editing (grammar and style). All content was reviewed, revised, and approved by the authors, who assume full responsibility for the manuscript.

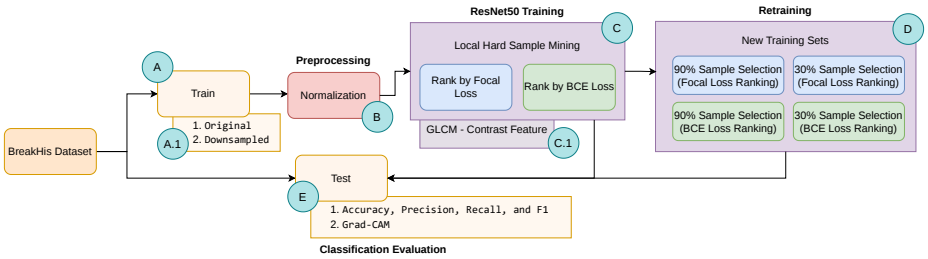

Figure 1. The proposed methodology representation.

microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). BreakHis comprises 2,480 benign samples and 5,429 malignant samples, each measuring 700 x 460 pixels. For this study, we selected images at a magnification of 400X for a binary classification task, in which samples were categorized as either "benign" or "malignant". We used only folds 1 and 2 from the original 5-fold structure of the initial study [Spanhol et al. 2016] for comparison due to resource limitations. Each fold contains the entire dataset, with differences in how samples are organized for training and testing. All dataset instances are present in each fold, but the training and testing partitions are rearranged according to the cross-validation scheme. (Fig. 1-A). Due to class imbalance, two different training subsets were created, one that contained all original samples from each fold, and one in which random downsample is applied to balance both labels (Fig. 1-A.1). All samples were normalized before model training, and no data augmentation was applied, aiming to analyze the impact of HSM techniques on the raw data (Fig. 1-B).

Local Hard Sample Mining. To identify the most challenging samples, we trained a ResNet50 architecture, since this architecture obtains minimal performance losses with data pruning according to [Liu et al. 2025b]. During the training process, we calculated both Focal Loss (FL), developed by [Lin et al. 2017], and Binary Cross-Entropy (BCE) Loss for each sample. These loss functions were used to compare sample selection based on loss ranking, considering FL's adaptability to unbalanced data and its capacity to emphasize the importance of hard samples [Liu et al. 2025a] (Fig. 1-C). The ranked images were compared using the contrast feature from the Gray Level Co-occurrence Matrix (GLCM) to analyze textural relationships between pixels in complex samples. GLCM is commonly used in histopathological images to detect tissue micro-architecture and identify subtle abnormalities [Şaban Öztürk and Akdemir 2018, Durgamahanthi et al. 2021] (Fig. 1-C.1).

Retrainig. Following [Liu et al. 2025b], who established 90% and 30% as the upper and lower bounds for sample selection rates, we trained new models using the top 90% and top 30% of the most challenging samples from the original training set, ranked by each loss function (Fig. 1-D). The testing set was consistent across all models, using the original testing sets provided by the creators of the BreakHis dataset.

Evaluation. To evaluate the effects of sample pre-selection in our classification task, we calculated Accuracy (Acc), Precision (Pr), Recall (R) and F1-score (F1) for each training setting. Additionally, to investigate the model's behavior, we generated Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps to visualize its learning

(Fig. 1-E). Grad-CAM helps to identify specific patterns within the input image by analyzing the gradients of the most significant regions based on the feature maps from the last convolutional layer of the trained model [Selvaraju et al. 2017, Peta and Koppu 2024].

4. Results and Discussion

This section presents results from sample pre-selection using BCE and FL ranking in breast cancer binary classification. Experiments were conducted on a server with an NVIDIA A100 GPU, utilizing Python 3.12.10 and PyTorch 2.7.0+cu126. We trained the models for 20 epochs with a batch size of 32, using the Adam optimizer, sigmoid activation, and a learning rate of $1e^{-4}$.

Regarding RQ1. To analyze the effects of HSM on training sample pre-selection, the ResNet50 models were initially trained using the original training data for comparison purposes. Table 1 presents the results of a baseline ResNet50 training, which utilized the original data without HSM. Table 2 summarizes the metrics used to evaluate HSM while training new ResNet50 models with selection rates ranging between 30% and 90%. The "Fold" column indicates the specific fold from the original 5-fold organization of the BreakHis dataset. The "Unbalanced or Balanced" column denotes if the model is trained with the original unbalanced dataset ("U") or with the balanced data ("B") from random downsampling. The arrows next to the metrics indicate whether the ideal value is closer to 0 (↓) or 1 (↑). The best metrics for each fold are highlighted in bold based on Acc. In the HSM experiments, the "Selection Rate" column indicates the proportion of ranked samples used to train new models, while the "Loss" column displays the loss function applied to rank the training samples. We opted not to summarize the metrics of both training folds by their mean because each fold had its own testing set in the original 5-fold experiments [Spanhol et al. 2016].

The initial experiments with the original data (Table 1) indicate that both balanced versions for Fold 1 and Fold 2 reduce training time with minor performance variations. For Fold 1, the unbalanced version achieved better classification performance, while the balanced version of Fold 2 delivered superior results with a variation of less than 2% in Acc for both unbalanced and balanced cases. When comparing the baseline and pruning experiments (Table 1 and Table 2), Pr is preserved with minor variations across pruned sets even when a more aggressive selection is applied (e.g., 0.9240 → 0.8984, decrease of 2.77%, between baseline U1 training set and U1 with a selection rate of 30% and FL ranking loss). Some cases of data pruning present a promising increase of R (e.g., 0.8857 → 0.9247, increase of 4.4%, between baseline U2 training set and U2 with a selection rate of 30% and BCE ranking loss), indicating that HSM can help boost recall in the context of medical images with high tissue complexity between samples. Pr and R reflect on F1, which remains stable across all models, indicating that pruning down to 30% suggests negligible overall performance cost. Training time is the most compelling efficiency gain, cutting nearly half of training time at 30% selection (e.g., 9.48 minutes → 5.17 minutes, decrease of 45.46%, between baseline U1 training set and U1 with a selection rate of 30% and BCE ranking loss). These results demonstrate that models trained with HSM for data pre-selection can achieve similar classification performance with only minor decreases, using up to half the training time and only 30% of the original training data.

To further analyze the impact of sample selection in model training, Figure 2

Table 1. Baseline performances of ResNet50 models trained with the original unbalanced dataset, and it's downsampled balanced counterpart.

Fold	Unbalanced or Balanced	Acc ↑	Pr ↑	R ↑	F1 ↑	Time (min) ↓
1	U1	**0.8901**	**0.9240**	**0.9019**	**0.9128**	9.48
	B1	0.8779	0.9005	0.9048	0.9091	**7.10**
2	U2	0.8213	0.8611	0.8857	0.8732	9.84
	B2	**0.8375**	**0.9251**	**0.8338**	**0.8770**	**7.12**

Table 2. Performance of ResNet50 on BreakHis dataset with different ranking loss functions and sample selection rates.

Fold	Unbalanced or Balanced	Selection Rate	Loss	Acc ↑	Pr ↑	R ↑	F1 ↑	Time (min) ↓
1	U1	90%	BCE	**0.8916**	**0.9141**	**0.9163**	**0.9152**	8.67
			FL	0.8794	0.8861	0.9306	0.9078	8.33
		30%	BCE	0.8824	0.9089	0.9067	0.9078	**5.17**
			FL	0.8885	0.8984	0.9306	0.9142	5.22
	B1	90%	BCE	0.8733	0.8798	0.9282	0.9034	8.88
			FL	**0.8779**	**0.8930**	**0.9187**	**0.9057**	9.31
		30%	BCE	0.8351	0.8656	0.8780	0.8717	5.65
			FL	0.8687	0.8915	0.9043	0.8979	**5.42**
2	U2	90%	BCE	0.8357	0.9038	0.8545	0.8785	8.84
			FL	**0.8412**	**0.9068**	**0.8597**	**0.8827**	9.34
		30%	BCE	0.8394	0.8558	0.9247	0.8889	**4.81**
			FL	0.8285	0.8699	0.8857	0.8777	4.85
	B2	90%	BCE	**0.8484**	**0.9239**	**0.8519**	**0.8865**	6.75
			FL	0.8339	0.9150	0.8390	0.8753	6.70
		30%	BCE	0.8267	0.8292	0.9455	0.8835	**4.05**
			FL	0.8249	0.8318	0.9377	0.8816	4.06

shows a sample correctly classified by a model trained with the original training set and a model trained with a selection rate of 90%, using BCE as the ranking loss function for Fold 1. Despite the pruning, the heatmap still emphasizes the same tissue region. This illustrates that a model trained on pre-selected samples exhibits learning behavior similar to a model trained on the raw dataset.

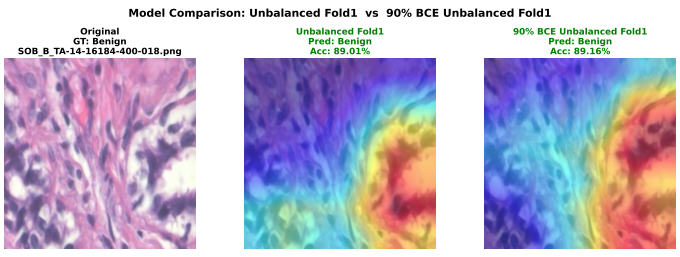

Figure 2. Grad-cam heatmap to compare model learning between models trained with the complete and pruned training sets for Fold 1.

Regarding RQ2. To analyse hard samples in a context of histology images classification, the contrast feature of GLCM was used. Figure 3 illustrates the distribution of the GLCM contrast feature across the original and pruned datasets for the unbalanced training set in Fold 1, using BCE and FL as the ranking loss function. To verify changes in contrast distributions, we applied the Mann-Whitney U test between benign and malignant labels and reported the p-value. In the original dataset, the p-value indicates clear difference in contrast between benign and malignant samples. As the selection rate decreases to 90%, the p-value begins to increase, suggesting a narrowing gap between hard and easy samples in terms of contrast. Notably, at 30% selection with FL ranking, the contrast distributions of benign and malignant samples overlap, with the p-value indicating no statistically significant distinction between classes. This progressive loss of contrast separability suggests that samples with ambiguous contrast are consistently ranked among the most difficult ones, and that models trained on the top 30% most ambiguous samples can achieve classification performance similar to that of models trained on the original training set in less time. It also points to FL as a more promising loss for sample ranking based on textural characteristics, as it emphasizes samples near the decision boundary, whereas BCE primarily reflects overall prediction error and can be influenced by outliers.

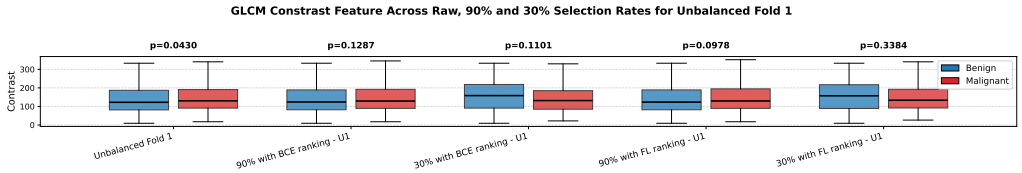

Figure 3. Boxplots for the GLCM contrast feature across original (raw) dataset, and selection rates of 90% and 30% based on BCE and FL ranking.

5. Conclusion and Future Works

This work presented an experimental approach to evaluate the impact of sample selection with a local Hard Sample Mining (HSM) method on the pruning of medical imaging datasets using two different loss functions during model training. The experiments show that pruning the training set based on the sample complexity ranked by the training loss function achieved promising classification results and training time. Initial analysis on ranked images show that samples with ambiguous GLCM contrast between benign and malignant histology images are considered complex samples in this classification context based on Focal Loss ranking. As future work, we plan to extend the experiments for a multiclass classification task on all five folds of the BreakHis dataset. Data Augmentation techniques will also be employed to achieve the best possible trade-off between sample distribution and model training efficiency.

References

Chen, K., Sun, S., Zhao, J., Wang, F., and Zhang, Q. (2025). Multi-view representation for pathological image classification via contrastive learning. *International Journal of Machine Learning and Cybernetics*, 16(4):2285–2296.

Chengxiao, Y., Xiaoyang, Z., Ahmed, A., Tunio, M. H., and Shuhuan, F. (2023). Diffusion-enhanced magnified histopathological images for robust breast cancer clas-

sification. In *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 1–7. IEEE.

Durgamahanthi, V., Anita Christaline, J., and Shirly Edward, A. (2021). Glcm and glrlm based texture analysis: Application to brain cancer diagnosis using histopathology images. In Dash, S. S., Das, S., and Panigrahi, B. K., editors, *Intelligent Computing and Applications*, page 691–706, Singapore. Springer.

Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, 32(4):582–596.

Kumar, T., Brennan, R., Mileo, A., and Bendechache, M. (2024). Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*, 12:187536–187571.

Li, Y., Yu, Y., Zou, Y., Xiang, T., and Li, X. (2022). Online easy example mining for weakly-supervised gland segmentation from histology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–587. Springer.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Liu, L., Liang, Y., Yan, X., Huangfu, L., Samtani, S., Yu, Z., Zhang, Y., and Zeng, D. D. (2025a). Hard sample mining: A new paradigm of efficient and robust model training. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–21.

Liu, L., Zhang, P., Liang, Y., Liu, J., Morra, L., Guo, B., Yu, Z., Zhang, Y., and Zeng, D. D. (2025b). -Razor: Hardness-Aware Dataset Pruning for Efficient Neural Network Training. *IEEE Transactions on Computational Social Systems*, 12(3):957–971.

Peta, J. and Koppu, S. (2024). Explainable soft attentive efficientnet for breast cancer classification in histopathological images. *Biomedical Signal Processing and Control*, 90:105828.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462.

Xue, C., Dou, Q., Shi, X., Chen, H., and Heng, P.-A. (2019). Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, page 1280–1283.

Şaban Öztürk and Akdemir, B. (2018). Application of feature extraction and classification methods for histopathological image using glcm, lbp, lbglcm, glrlm and sfta. *Procedia Computer Science*, 132:40–46. International Conference on Computational Intelligence and Data Science.