Addressing Class Imbalance in Renal Amyloidosis Classification: A Comparative Study of Few-Shot Learning and Conventional Machine Learning Techniques

Alexsandro Silva Santos[1][0009−0005−8934−5009], Luciano Rebouças de Oliveira[2][0000−0001−7183−8853], Washington Luis Conrado dos Santos[3][0000−0002−5075−1254], and Angelo Amancio Duarte[4][0000−0001−7446−1342]

[1] Graduate Program in Computer Science, State University of Feira de Santana, Av. Transnordestina, s/n, 44036-900. Feira de Santana, Brazil
alexsandro.ssantos75@gmail.com

[2] Intelligent Vision Research Lab, Computer Science Department, Federal University of Bahia, Rua Prof. Aristides Novis, 2, 40210-630. Salvador, Brazil
lrebouca@ufba.br

[3] Structural and Molecular Pathology Lab, Gonçalo Moniz Institute, Fundação Oswaldo Cruz, Rua Waldemar Falcão, 121, 40296-710. Salvador, Brazil
washington.santos@fiocruz.br

[4] High-Performance Computing Lab, Department of Technology, State University of Feira de Santana, Av. Transnordestina, s/n, 44036-900. Feira de Santana, Brazil
angeloduarte@uefs.br

Abstract. Class imbalance presents a significant challenge in Computational Pathology, particularly in classifying rare diseases such as renal amyloidosis. This paper investigates the effectiveness of Few-Shot Learning (FSL), specifically through prototypical networks, alongside conventional methods to enhance the automatic classification of renal glomeruli from biopsy images. A novel multi-stain dataset is introduced, comprising 11,674 annotated images across nine glomerular lesion classes, including amyloidosis, stained with four different dyes. The study compared baseline CNN models with FSL approaches, both with and without Cost-Sensitive Learning (CSL). The FSL-CSL-Ensemble achieved the highest F1-Score of 93.8%, surpassing the performance of related studies that addressed datasets with less severe imbalance ratios. This study underscores the potential of FSL in classifying renal amyloidosis, especially when combined with CSL, and suggests the possibility of eliminating the need for Congo red staining, the current gold standard for diagnosis. The findings highlight the necessity of developing innovative approaches like FSL to improve outcomes in medical image analysis, where data scarcity is prevalent.

Keywords: Class Imbalance · Few-shot Learning · Computational Pathology · Glomeruli classification.

1 Introduction

Computational Pathology (CPATH) has emerged as a powerful tool in the diagnosis and classification of various diseases, particularly in the field of nephrology. Among the myriad of renal pathologies, the ones caused by glomerular lesions present a unique challenge due to the diverse morphological characteristics and clinical implications. Accurate classification of glomerular lesions is crucial for proper diagnosis, prognosis, and treatment planning. However, the complexity and variability of glomerular lesions, coupled with the scarcity of certain pathologies, that often yield imbalanced classes, pose significant challenges for traditional machine learning approaches in CPATH.

Class imbalance is a pervasive issue in CPATH, particularly in the context of disease classification. This imbalance occurs when one or more classes in a dataset are significantly underrepresented compared to others. In the context of glomerular lesions, this issue is especially pronounced due to factors inherent to medical data and disease prevalence, such as the rarity of certain diseases or conditions, difficulty and cost of data acquisition, ethical and privacy concerns, variability in disease progression, and bias in data collection.

A prime example of this class imbalance challenge is evident in the classification of renal amyloidosis, a rare but severe condition affecting the kidneys. Amyloidosis is characterized by the abnormal deposition of misfolded proteins, called amyloid fibrils, in various organs, including the kidneys. In renal amyloidosis, these deposits accumulate in the *glomeruli*, the kidney's filtration units, leading to progressive organ dysfunction and potentially fatal outcomes [15].

The classification of renal amyloidosis presents several unique challenges: a) Rarity- Amyloidosis is a relatively uncommon condition, resulting in limited available data for training machine learning models; b) Morphological Similarity- Amyloid deposits can sometimes resemble other glomerular lesions, making differentiation challenging even for experienced pathologists; c) Staining Variability- The gold standard for amyloidosis diagnosis, Congo red staining [24], can be inconsistent and requires specialized expertise to interpret correctly; and d) Heterogeneity- Amyloidosis can present with various patterns and distributions of amyloid deposits, further complicating classification efforts.

These challenges contribute significantly to the under-representation of amyloidosis in datasets with images of glomeruli. Consequently, in the context of computational pathology, samples of renal amyloidosis are vastly outnumbered by those of more common renal conditions or healthy tissue. This pronounced imbalance poses a substantial challenge for traditional machine learning approaches, which tend to be biased towards the majority class, potentially leading to missed diagnoses of this critical condition.

Classical approaches to tackle class imbalance typically fall into three categories: data-level methods, algorithm-level methods, and hybrid methods [7]. Data-level techniques, such as oversampling and undersampling, aim to balance the dataset by adjusting the number of samples in each class. Algorithm-level methods, including cost-sensitive learning and ensemble techniques like bagging and boosting, modify the learning process to compensate for class imbalance. Hy-

brid methods combine both, data and algorithm-level approaches. While these techniques have shown success in many applications, they often struggle with extreme imbalances or when dealing with limited data availability, as is common in rare diseases like renal amyloidosis.

Few-Shot Learning (FSL) approaches [5] have emerged as a promising solution to address data sparsity scenarios and mitigate the operational costs associated with dataset annotation, particularly in contexts where data is limited [23]. FSL is designed to learn effectively from a small number of labeled examples, making it especially suitable for rare disease classification tasks in computational pathology.

While initially developed to tackle data scarcity, FSL has increasingly found application in addressing class imbalance problems [2]. This expansion of FSL's utility has given rise to a new research direction known as Class Imbalance Few-Shot Learning (CIFSL) [13]. CIFSL leverages the inherent ability of FSL to learn effectively from a small number of labeled examples, making it especially suitable for rare disease classification tasks, offering the potential to overcome the limitations of traditional class imbalance techniques, especially when dealing with rare conditions in computational pathology. However, the application of CIFSL to the specific challenge of glomerular lesion classification, particularly renal amyloidosis, remains largely unexplored.

In this study, we present a comparative analysis between conventional techniques and Few-Shot Learning (FSL) as strategies for addressing class imbalance in glomerular disease diagnostics, with a specific focus on amyloidosis. Our approach begins with the development of a baseline model using established methods to combat class imbalance, namely resampling [21] and cost-sensitive learning (CSL) [12], which serves as a benchmark for subsequent comparisons. We then construct a series of FSL classifiers utilizing prototypical networks, employing the same architectural foundations as our baseline models for the embedding functions. To further enhance performance, we integrate standard N-way-K-shot episodic training with CSL. Finally, we create an ensemble-based model using the top-performing individual models. Throughout our analysis, we employ the F1-Score as our primary metric for model comparison. This choice is deliberate, as the F1-Score provides a balanced measure of precision and recall, avoiding the potential bias towards the majority class that can occur with other metrics such as accuracy [11]. This comprehensive approach allowed us to rigorously evaluate the efficacy of FSL in managing class imbalance within the context of rare glomerular disease classification.

This study aimed to address critical gaps in computational pathology for renal disease classification through several key objectives. We introduce a novel, large-scale dataset of 11,674 annotated glomeruli images across nine lesion classes and four staining techniques, addressing the scarcity of comprehensive, multi-stain datasets in renal pathology (dataset available under request). Our research compares the performance of conventional machine learning techniques with Few-Shot Learning (FSL) approaches in classifying renal amyloidosis, focusing particularly on addressing extreme class imbalance. We explore an innovative

combination of FSL with Cost-Sensitive Learning to further improve classification performance in the context of severe class imbalance. Of particular interest is our investigation into the use of non-specific stains for amyloidosis classification, which could potentially eliminate the need for specialized Congo red staining in the diagnostic process. This aspect is especially significant as the current method of diagnosing amyloidosis requires biopsy slides to be processed using Congo red, a specific staining technique. Our proposed method aims to accurately diagnose amyloidosis from glomeruli images stained with common dyes like Periodic Acid-Schiff (PAS) and Hematoxylin and Eosin (H&E). Our results showed that this approach could significantly simplify the diagnostic process, leading to faster and more cost-effective diagnoses of this disease.

By addressing these objectives, our study seeks to advance the field of computational pathology in renal disease classification, potentially improving diagnostic accuracy and efficiency in clinical practice. Furthermore, our findings may serve as a basis for the development of methods for the classification of other rare diseases characterized by significant class imbalance in medical imaging datasets.

2 Related Works

In addressing the class imbalance in medical imaging, researchers have employed various innovative approaches. Mahbub *et al.* proposed an algorithmic method using a novel cost function, Center-Focused Affinity Loss (CFAL), for histological dataset imbalance. They achieved an F1-Score of up to 83% on a substantial dataset of 277,524 samples with an imbalance ratio (IR) of approximately 1:3 [9]. Walsh and Tardy focused on mammography image classification, comparing traditional imbalance techniques with generative adversarial networks (GAN). Their proposed method, "Artifacting," achieved an AUCROC of 76.8% on a highly imbalanced dataset (IR 1:19) containing 20.000 images [22]. Raj *et al.* introduced a data augmentation technique called the "Crossover-based Technique," which generates new samples by combining existing images. Applied to CNN training on three medical datasets, this method achieved an impressive Macro F1-Score of 98% in a multi-class brain tumor detection task [14]. These studies demonstrate the potential of diverse approaches in addressing class imbalance across various medical imaging domains, from histopathology to mammography and brain tumor detection, highlighting the ongoing challenge and the need for innovative solutions in this field.

Few-shot learning (FSL) has gained attention not only for addressing data scarcity but also for tackling class imbalance, sometimes referred to as Class Imbalance Few-shot Learning (CIFSL). Deng and Li combined Transfer Learning (TL), FSL, resampling techniques, and image masking methods for white blood cell classification and counting in blood samples with an imbalance ratio of 1:6 [3], achieving an AUCROC of up to 88%. Medela et al. applied an FSL approach with Siamese Neural Networks to transfer knowledge from a multiclass colon dataset to healthy and cancerous tissues of the colon, breast, and lung [10]. Using only 60 samples per class in the support set, they achieved up to 90% Bal-

anced Accuracy (BAC) on balanced datasets. Abbas proposed the Intelligence Medical Imaging Recognition (IMR-FSL) model for image retrieval, testing it on the TCIA (Clark et al., 2013) and KVASIR (Pogorelov et al., 2017) datasets [1]. Their model demonstrated impressive performance with 95% sensitivity, 96.5% specificity, 0.96 AUC, and 97.5% accuracy. Titoriya and Singh utilized Prototypical Networks (PN) and Model Agnostic Meta-Learning (MAML) across four datasets for cancer diagnosis, two of which were imbalanced [20]. Their study reported accuracy up to 84.56% for a 2-way-2-shot configuration on the query set.

These studies showcase the versatility of FSL in addressing both data scarcity and class imbalance across various medical imaging applications, from blood cell analysis to cancer diagnosis.

3 Methods

Our methodology aimed at comparing baseline classifier models trained on the entire dataset of images against models trained solely on images stained with the Periodic Acid-Schiff (PAS) dye. This comparison allowed us to assess the impact of stain selection on model performance. Additionally, we investigated the potential benefits of incorporating Cost-Sensitive Learning (CSL) to address the inherent class imbalance within the dataset, evaluating whether this approach can enhance the classifiers' performance. Furthermore, we aimed to explore the efficacy of Few-Shot Learning (FSL) techniques to determine if they can further improve model accuracy when dealing with limited data, as is often the case in medical imaging. Through these evaluations, our goal was to identify optimal strategies for enhancing the automatic classification of renal amyloidosis.

Throughout our analysis, we employ the F1-Score as our primary metric for model comparison. This choice is deliberate, as the F1-Score provides a balanced measure of precision and recall, making it particularly suitable for evaluating classifiers on imbalanced datasets [4].

In the context of our highly imbalanced dataset, where amyloidosis cases are significantly outnumbered by other glomerular lesions, accuracy alone can be misleading. A model that simply predicts the majority class for all instances could achieve high accuracy without actually identifying any amyloidosis cases. The F1-Score, being the harmonic mean of precision and recall, penalizes such behavior and provides a more nuanced evaluation of model performance.

Moreover, in clinical applications like renal pathology, both false positives and false negatives can have significant consequences. False positives may lead to unnecessary treatments or anxiety for patients, while false negatives could result in delayed or missed diagnoses. The F1-Score, by considering both precision and recall, helps us balance these concerns and identify models that perform well in both aspects.

Additionally, the F1-Score is particularly useful when dealing with rare conditions like amyloidosis. It gives equal weight to precision and recall, ensuring that models are evaluated not just on their ability to avoid false positives (preci-

sion), but also on their capacity to identify true positive cases (recall) in a sparse dataset.

By consistently using the F1-Score across our different experimental setups - from baseline models to Few-Shot Learning approaches - we maintain a standardized basis for comparison. This allows us to effectively assess the relative strengths of different methodologies in addressing the class imbalance challenge in glomerular lesion classification.

3.1 Dataset

We developed a comprehensive dataset comprising 11,674 glomeruli images, meticulously extracted from whole slide images (WSI) of renal biopsies. These high-quality images were stored in JPEG format to balance detail preservation and storage efficiency. Our dataset encompasses a wide spectrum of pathological conditions, reflecting the complexity of renal pathology: normal glomeruli (without lesions), amyloidosis lesions, pure sclerosis without crescent, hypercellularity-type lesions, pure hypercellularity-type lesions without crescent, crescentic glomerulonephritis, membranous nephropathy, sclerosis, and podocytopathy.

To ensure a robust representation of each condition, we employed four distinct histological staining techniques, each offering unique insights into glomerular structure and pathology: AZAN trichrome (AZAN), hematoxylin and eosin (HE), periodic acid-methenamine silver (PAMS), and periodic acid-Schiff (PAS).

The images in the dataset exhibit dimensional variability, ranging from 607 $\times$ 751 pixels to 1024 $\times$ 768 pixels, reflecting the natural variation in glomerular size and shape. All images maintain a consistent high resolution of 300 $\times$ 300 dpi, ensuring detailed visualization of glomerular structures.

Table 1 provides a detailed distribution of samples across pathological conditions and staining techniques, offering a quantitative overview of the dataset's composition. Figure 1 presents representative images of glomeruli stained with each technique, visually demonstrating the morphological diversity captured in our dataset. This comprehensive approach to dataset construction ensures a rich, varied foundation for training and evaluating our machine-learning models in glomerular lesion classification.

3.2 Data Pre-processing

The amyloidosis lesion was designated as the positive class, with all other glomerular lesions grouped into a single negative class. This resulted in a highly imbalanced dataset with an imbalance ratio of approximately 1:30, reflecting the rarity of amyloidosis in clinical settings.

For a complete model evaluation, we employed a 75:25 train-test split. The 75% training portion underwent K-fold cross-validation, so we could take a robust performance estimation by detecting variance in the model performance across different samples. This approach facilitated fair comparisons between different models. We partitioned the source set into $K = 5$ folds, where $(K-1)/K$

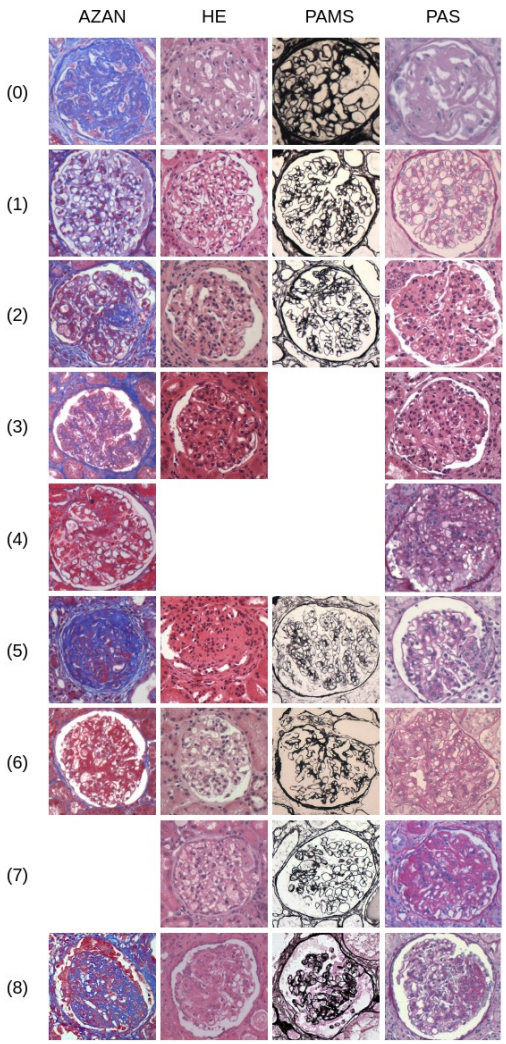

Fig. 1. Samples of images in the dataset according to the class of lesion and type of stain. Blank squares indicate no sample of a lesion in the stain. (0) Amyloidosis, (1) Normal, (2) Pure Sclerosis, (3) Hypercellularity, (4) Pure Hypercellularity, (5) Crescent Glomerulonephritis, (6) Membranous Nephropathy, (7) Sclerosis, (8) Podocytopathy

of the images were used for training and $1/K$ reserved for validation. The decision to use five folds instead of the more commonly recommended ten folds in cross-validation literature was driven by the limited number of samples in some classes, particularly the rare lesions such as amyloidosis. With five folds, we ensured a sufficient number of minority class samples in each fold for meaningful evaluation, while still maintaining the benefits of cross-validation. This strat-

Table 1. Original dataset distribution, by lesion and staining. (0) Amyloidosis, (1) Normal, (2) Pure Sclerosis, (3) Hypercellularity, (4) Pure Hypercellularity, (5) Crescent Glomerulonephritis, (6) Membranous Nephropathy, (7) Sclerosis, (8) Podocytopathy.

Class		AZAN	HE	PAMS	PAS	**Total**
pos	0	31	145	96	102	**374**
neg	1	223	1,585	345	542	**2,695**
	2	234	672	104	472	**1,482**
	3	257	1.890	0	987	**3,134**
	4	60	0	0	164	**224**
	5	121	467	157	359	**1,104**
	6	136	712	324	367	**1,539**
	7	0	276	122	219	**617**
	8	90	65	106	244	**505**
	Total	**1,152**	**5,812**	**1,254**	**3,456**	**11,674**

egy struck a balance between robust performance estimation and the practical constraints imposed by our dataset's class imbalance, allowing for more reliable model assessment in the context of rare disease classification.

Our study encompassed two distinct experimental paradigms: a comprehensive multi-stain analysis utilizing the entire dataset of 11,674 images across all four staining techniques, and a single-stain-focused analysis employing a subset of 3,456 images exclusively stained with Periodic Acid-Schiff (PAS). Both experimental sets underwent identical preprocessing and partitioning procedures to ensure consistency and comparability of the results. The selection of PAS for our focused analysis was based on several key factors. Firstly, PAS uniquely captured all types of lesions present in the original dataset. Secondly, PAS is ubiquitous in nephropathology and general pathology practices worldwide. Its pervasiveness is due to its ability to highlight important structural elements such as basement membranes, glycogen, and neutral mucopolysaccharides, making it an indispensable tool in the diagnosis of various renal pathologies. This widespread use enhances the translational potential and clinical relevance of our findings. Lastly, focusing on PAS offered the potential to streamline diagnostic workflows. This experiment was designed to evaluate whether PAS-only models could accurately classify the full spectrum of glomerular lesions, compare their performance to multi-stain models, and explore the potential for optimizing biopsy analysis. By assessing the efficacy of this single, commonly used stain for comprehensive lesion classification, we aimed to investigate opportunities to reduce procedural complexity, improve diagnostic efficiency, and enhance cost-effectiveness in pathological examinations. This approach not only addresses the technical aspects of machine learning in pathology but also considers the practical implications for clinical workflow and resource allocation in diagnostic nephropathology, potentially paving the way for more standardized and efficient diagnostic processes.

To establish a robust comparison, we implemented multiple strategies to address the inherent class imbalance in our dataset. Initially, we established a base-

line training the models with a dataset using classical random oversampling to achieve a balanced 1:1 ratio of positive (amyloidosis) to negative (other lesions) samples across all folds, providing a standard benchmark in imbalanced learning scenarios. Then, we took the unbalanced dataset and developed models incorporating cost-sensitive learning, assigning higher weights to the minority class during training to mitigate bias towards the majority class without altering the underlying data distribution. For the Few-Shot Learning (FSL) experiments utilizing prototypical networks, we deliberately avoided resampling techniques. This decision was based on the fundamental principle of prototypical networks, which leverage the average of sample embeddings to calculate class prototypes. We hypothesized that introducing duplicate samples through oversampling might introduce noise and potentially degrade the quality of the learned prototypes, leading to suboptimal model performance.

To ensure the validity and generalizability of our results, we conducted final model validation on the initially reserved 25% test set, which crucially maintained the original dataset's imbalanced proportions, reflecting real-world class distributions and providing a more realistic assessment of model performance in clinical scenarios. This comprehensive methodological approach enables a rigorous and statistically sound evaluation of each model's predictive capabilities across diverse data subsets, allows for a fair comparison between traditional machine learning approaches and FSL techniques in the context of extreme class imbalance, mitigates potential biases, reduces the risk of overfitting, and provides insights into the effectiveness of different strategies for handling class imbalance in the specific context of glomerular lesion classification.

3.3 CNN models

We employed six pre-trained convolutional neural network (CNN) architectures: EfficientNet-B0, EfficientNet-B4, Inception-v3, ResNet-18, ResNet-50, and VGG-16. These architectures were chosen based on their proven performance in various image classification tasks and their distinct architectural features.

EfficientNet models, developed by Tan and Le [19], are known for their ability to balance network depth, width, and resolution. They achieve state-of-the-art accuracy on ImageNet while being up to 10 times smaller and faster than previous models. The B0 and B4 variants were selected to compare the performance of a smaller, more efficient model (B0) against a larger, potentially more powerful one (B4) in the context of our glomerular lesion classification task.

Inception-v3, introduced by Szegedy et al. [18], is designed to be computationally efficient while maintaining high accuracy. Its use of factorized convolutions and aggressive regularization makes it particularly suitable for tasks where computational resources may be limited, as is often the case in medical image analysis settings.

Residual Networks (ResNet), developed by He et al.[6], address the vanishing gradient problem in deep networks through the use of skip connections. This allows for the training of much deeper networks, potentially capturing more complex features in the images. We included both ResNet-18 and ResNet-50 to

evaluate whether the increased depth of ResNet-50 provides significant benefits in our specific classification task.

Despite being an older architecture, VGG-16 [8] remains relevant due to its simplicity and effectiveness. Its uniform architecture makes it easier to interpret and modify, which can be advantageous when fine-tuning for specific medical imaging tasks.

The models were trained using a learning rate of 0.001, batch size of 32, and Stochastic Gradient Descent optimizer. Training proceeded for a maximum of 100 epochs, with an early stopping mechanism implemented with patience of 10 epochs to prevent overfitting. These diverse architectures were selected to provide a comprehensive baseline for our study. By comparing their performance, we aim to identify which architectural features are most beneficial for glomerular lesion classification, particularly in the context of the class imbalance present in our dataset. This comparison also allows us to assess whether more complex, modern architectures offer significant advantages over simpler, more established models in this specific medical imaging context. Furthermore, these pre-trained models allow us to leverage transfer learning, potentially mitigating the impact of our limited dataset size.

Furthermore, these pre-trained models allow us to leverage transfer learning, potentially mitigating the impact of our limited dataset size. While these models were initially trained on natural images (ImageNet), previous studies have shown that the low-level features learned by CNNs can be effectively transferred to medical imaging tasks, providing a strong starting point for our fine-tuning process.

3.4 Classical approach for class imbalance

To establish a comprehensive baseline for our study, we implemented two distinct strategies to address the inherent class imbalance in our dataset. First, we trained models using a dataset balanced through classical random oversampling. This technique achieved a 1:1 ratio of positive (amyloidosis) to negative (other lesions) samples across all folds, providing a standard benchmark in imbalanced learning scenarios. Second, we developed models using the original unbalanced dataset, incorporating cost-sensitive learning (CSL). This approach assigned weights to the loss function inversely proportional to each class's sample count, thereby prioritizing the minority class during training. This method mitigates bias towards the majority class without altering the underlying data distribution.

To rigorously assess the impact of CSL, we trained each architecture both with and without its implementation. The models were subsequently ranked in descending order based on their F1 scores across the 5-fold cross-validation, with the F1 score chosen due to its balanced consideration of precision and recall, crucial in imbalanced classification tasks.

Following this initial evaluation, we identified the three top-performing architectures based on their aggregate scores. These selected models underwent full training using the entire training set (75% of the original dataset), with the remaining 25% reserved as a hold-out set for final validation. This approach

maximized data diversity during training, enhancing the generalizability of our classifiers.

To leverage the collective strengths of these top-performing models, we developed an ensemble-based classifier, referred to as *Baseline-ensemble*. This ensemble method combines the predictions of individual models, potentially improving overall accuracy and robustness.

This systematic approach to model selection, training, and ensemble construction ensured a robust baseline for comparison with our Few-Shot Learning models. It provides a comprehensive evaluation of different architectural and training strategies in the context of glomerular lesion classification, particularly for the rare condition of renal amyloidosis.

3.5 Few-Shot Learning for class imbalance

Few-Shot Learning (FSL) leverages prior knowledge gained from training on a large, labeled dataset, to perform efficiently on small classification tasks within a specific domain of interest. FSL employs a unique training paradigm known as episodic training, which mimics the few-shot scenario during the learning process. In the FSL framework, the core concept is the N-way-K-shot task. Here, N represents the number of classes to be distinguished, while K denotes the number of examples provided for each class. These examples form the *support set*, a small, labeled dataset used for learning. The model's performance is then evaluated on a separate *query set*, which contains new, unseen examples from the same classes. This approach allows the model to adapt quickly to new tasks with minimal data [25].

Among FSL methodologies, metric-based approaches, particularly prototypical networks, have gained prominence. In these networks, the support set samples are used to generate class prototypes. This is achieved through an embedding function, typically a Convolutional Neural Network (CNN), denoted as f_ϕ, where ϕ represents the network parameters [16]. The function f_ϕ transforms input samples into a feature space where similar samples cluster together. Class prototypes are then computed as the mean of the embedded support samples for each class.

Classification of a new sample in prototypical networks involves comparing its embedding to these class prototypes. This comparison is performed using a distance function d, commonly either cosine similarity or Euclidean distance. The new sample is assigned to the class whose prototype is nearest in the embedding space, effectively leveraging the model's ability to learn meaningful representations from limited data [17].

For training the FSL models we used a metric-based approach with prototypical networks. Each model was trained using the same folds used in the baseline experiment, with each fold comprising a support set for prototype generation and a query set for validation. We implemented episodic training following the standard *N-way-K-shot* FSL paradigm, with $N = 2$ (binary classification) and $K = 30$ for the training stage's support set, and $K = 15$ for the validation stage's query set. The values for K were selected based on the number of samples in the minority class. Specifically, for each fold, the training set included 60

samples, and the validation set contained 16 minority samples, which facilitated testing on a final imbalanced set with 25 minority samples. The models were trained over 100 epochs, with 200 episodes per epoch, and included an early stopping mechanism that was triggered after 10 epochs without performance improvement.

During each fold's episodes, we stored the optimal parameter and prototype states for each model. We hypothesized that cross-validation would enhance prototype construction by leveraging the entire training set. Post-training, each model was evaluated against the reserved 25% imbalanced test set using its best fold-specific state.

We explored two distance functions for classification: Euclidean distance and cosine similarity. Experimental results demonstrated a significant performance advantage for cosine similarity.

Following a methodology analogous to the baseline classifier, we selected the top three performing models based on the F1 score to construct two ensemble-based models: *FSL-Ensemble* (without CSL) and *FSL-CSL-Ensemble* (with CSL applied). This approach aimed to leverage the collective predictive power of the most effective prototypical network models while mitigating class imbalance challenges.

4 RESULTS AND DISCUSSION

Here we present the results of the experiments outlined in Section 3, which focused on comparing the effectiveness of conventional machine learning techniques with Few-Shot Learning (FSL) for classifying amyloidosis in renal glomeruli. The experiments investigated the performance of pre-trained convolutional neural network (CNN) architectures as baseline models, both with and without cost-sensitive learning (CSL), and compared these to FSL models using prototypical networks. We explore the impact of stain selection on model performance, comparing models trained on the entire dataset (with all stains) to those trained exclusively on Periodic Acid-Schiff (PAS) stained images. Additionally, we discuss the effectiveness of combining conventional approaches like CSL with FSL techniques to enhance classification accuracy.

4.1 Results using classical approach

Figure 2 presents a comparative analysis of F1-scores achieved by various convolutional neural network (CNN) architectures in the classification of renal amyloidosis, with a particular emphasis on the impact of stain selection. The results demonstrate a consistent performance advantage for models trained on the complete dataset, which incorporates all four staining techniques, over those trained exclusively on Periodic Acid-Schiff (PAS) stained images. This performance disparity suggests that the diverse visual information provided by multiple staining methods contributes significantly to enhanced model accuracy.

Within each stain selection group, certain architectures exhibit superior performance. For the full dataset, VGG-16, ResNet-18, and EfficientNet-B0 achieve the highest F1 scores. In contrast, when trained solely on PAS-stained images, EfficientNet-B0, ResNet-50, and ResNet-18 emerge as the top performers. This variation in architectural efficacy across staining subsets indicates that the optimal choice of CNN architecture for amyloidosis classification may be contingent on the specific staining technique employed in the training data. These findings underscore the importance of considering both architectural design and staining methodology in developing robust classification models for renal pathology.

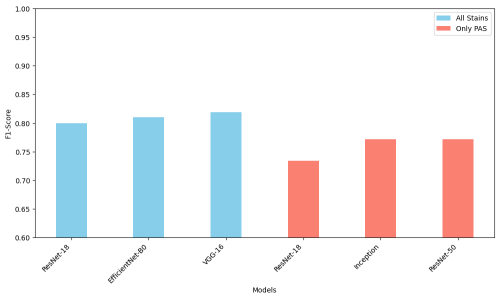

Fig. 2. Comparison of F1-Scores between models trained on samples from all stains (blue) and those trained solely on PAS-stained samples (red).

Based on the superior performance of models trained on samples from all stains, we selected these models for further optimization using cost-sensitive learning (CSL). Figure 3 illustrates the impact of the CSL application, which, as hypothesized, resulted in significant performance enhancements across all architectures. Notably, the ensemble classifier, integrating VGG-16, ResNet-18, and EfficientNet-B0, achieved the highest F1-Score among all models. This marked improvement can be attributed to the ensemble method's capacity to synergistically leverage the unique strengths of each constituent architecture, thereby enhancing overall classification accuracy and model robustness. The combination of diverse staining information, cost-sensitive learning, and ensemble techniques demonstrates a powerful approach to addressing the challenges of renal amyloidosis classification in imbalanced datasets.

4.2 Results using Few-Shot Learning

Figure 4 reveals a significant divergence in performance trends between baseline models and Few-Shot Learning (FSL) models. While baseline models trained on the complete multi-stain dataset consistently outperformed those trained exclusively on PAS-stained images (as shown in Figure 2), FSL models exhibit the opposite behavior. FSL models trained solely on PAS-stained images achieve

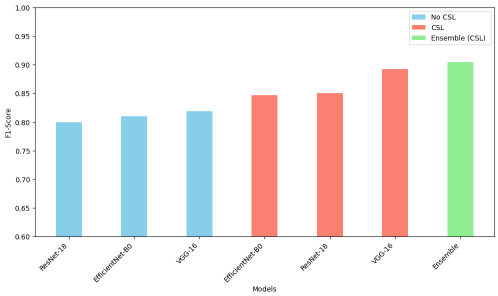

Fig. 3. Comparison of F1-scores for baseline models trained on multi-stain samples, without (blue) and with (red) cost-sensitive learning (CSL). The ensemble-based model (green) demonstrates superior performance.

superior results compared to their counterparts trained on the full multi-stain dataset.

This performance disparity can be attributed to the unique learning mechanism of prototypical networks employed in our FSL models. Prototypical networks generate class prototypes by averaging the embeddings of samples in the support set. When the training dataset encompasses images with diverse staining characteristics, as in the full dataset, the resulting embedding space can be heterogeneous. This heterogeneity potentially introduces outliers that negatively impact the representativeness of the prototypes.

Conversely, utilizing a dataset with a single stain, such as PAS, produces a more homogeneous embedding space, leading to more robust and representative prototypes. This homogeneity enables the FSL model to generalize more effectively from limited data, a crucial aspect of FSL where models are trained on a small number of samples.

Therefore, while a diverse dataset proves beneficial for traditional machine learning models, as evidenced by the baseline model performance, the unique characteristics of FSL and its reliance on prototype-based learning render a homogeneous dataset, even if limited to a single stain, more advantageous for achieving optimal performance. This finding underscores the importance of considering the specific learning paradigm when selecting and preparing datasets for different machine-learning approaches in medical image analysis.

The application of Cost-Sensitive Learning (CSL) to models trained on PAS-stained images yielded a significant performance improvement, as illustrated in Figure 5. Following the methodology employed for the baseline classifier, we constructed an ensemble using the three top-performing models. This ensemble achieved a remarkable F1-Score of 93.8%, surpassing all individual models. This outstanding result underscores the ensemble classifier's ability to effectively leverage the strengths of multiple architectures, particularly when combined with CSL in the context of Few-Shot Learning. The superior performance of this ap-

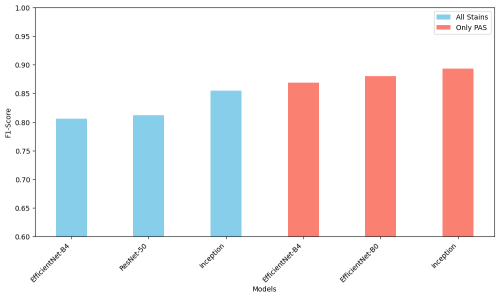

Fig. 4. Comparison of F1-scores for Few-shot Learning (FSL) models trained with images in all stains (blue) and models only trained with images in PAS stain (red). Models for PAS stain demonstrate superior performance.

proach demonstrates its potential for addressing the challenges of class imbalance and limited data in renal amyloidosis classification.

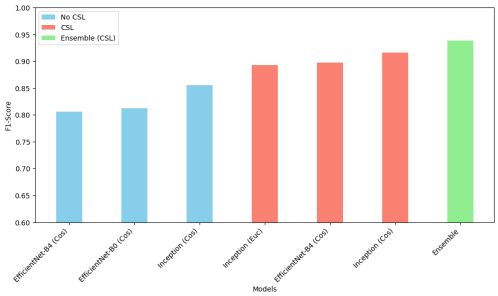

Fig. 5. Comparison of F1-scores for Few-Shot Learning (FSL) models trained on PAS-stained images, without (blue), and with (red) Cost-Sensitive Learning (CSL), and the ensemble model. Only one model yielded the best result using Euclidean distance (Euc). The vast majority achieve the best performance using cosine similarity (Cos) as distance metrics in the prototypical networks. The ensemble-based model (green) demonstrates superior performance.

Table 2 presents the F1 scores of the five best-performing classifiers. Few-shot learning (FSL) demonstrates significant potential for renal amyloidosis classification, particularly when combined with Cost-Sensitive Learning (CSL). However, several limitations warrant further investigation.

The generalizability of this approach to other glomerular lesions requires careful examination, as our study focused specifically on amyloidosis. Future research should evaluate FSL's efficacy in classifying lesions with diverse visual characteristics and varying levels of data availability. The reliance on PAS staining for optimal performance raises concerns about applicability in settings where

this technique is not standard practice. Our findings highlight the potential impact of specific staining methods on FSL model performance, emphasizing the need for comprehensive validation across various staining protocols.

While stratification helps mitigate bias in prototype construction, the risk of biased prototypes persists if the limited samples available do not fully capture the lesion's true diversity. Additionally, the computational demands and complexity of FSL, especially with sophisticated architectures like prototypical networks, may present implementation challenges in resource-constrained environments.

Our results indicate that conventional CNNs require substantially more data for effective generalization, as evidenced by the baseline performance disparity between models trained on the entire dataset versus those trained solely on PAS samples. In contrast, FSL classifiers, particularly prototypical networks, demonstrate superior performance with limited data. This phenomenon can be attributed to the fact that large volumes of data may introduce outliers, potentially compromising the accuracy of generated prototypes.

These findings underscore the potential of FSL in addressing the challenges of limited data and class imbalance in medical image classification, while also highlighting areas requiring further research and optimization.

Table 2. Rank of the 5 best classifiers. The term CSL indicates using of cost-sensitive learning while *Cos* and *Euc* indicate the type of distance metric for few-shot-learning (FSL) models.

Architecture	**Dataset**	**F1-Score**
FSL-CSL-Ensemble + CSL	PAS	0.938
Inception + CSL (Cos)	PAS	0.916
Baseline-Ensemble	ALL	0.905
EfficientNet-B4 + CSL (Cos)	PAS	0.897
Inception + CSL (Euc)	PAS	0.893

5 Conclusions

This study investigated the integration of conventional Machine Learning (ML) techniques with Few-Shot Learning (FSL) to improve the automatic classification of renal amyloidosis. The inherent imbalance in the dataset, with amyloidosis being a rare condition, posed significant challenges for traditional ML methods. The results indicate that while standard ML approaches, even those designed to address class imbalance, may not independently achieve robust performance, their combination with FSL shows considerable promise.

Our ensemble-based model achieved an impressive F1-score of 93.8%, surpassing related studies that dealt with datasets featuring less severe imbalance ratios. Incorporating established methods like Cost-Sensitive Learning (CSL) with FSL techniques significantly enhanced overall classification performance.

The superior outcomes observed with FSL models, particularly when applied to PAS-stained samples, highlight FSL's ability to leverage limited data for improved generalization, which is crucial in medical datasets often lacking data abundance.

A significant contribution of this research is the development of a novel multi-stain dataset consisting of 11,674 images of renal glomeruli (available under request), annotated across nine classes with four different stains. This dataset addresses a crucial gap in computational pathology, given the challenges associated with gathering and annotating such a large volume of glomerular images.

The study also highlights the potential for classifying amyloidosis without relying on Congo red staining, the current diagnostic gold standard. If successful, this innovative approach could significantly streamline the diagnostic process. The findings encourage further exploration of this methodology for other glomerular lesions, potentially leading to a computer-aided diagnosis tool that would greatly aid pathologists in diagnosing glomerular diseases.

Additionally, the research underscores the limitations of conventional CNN approaches when faced with limited data, as evidenced by lower performance with stratified data in baseline models. This emphasizes the need for innovative approaches like FSL to improve outcomes when dealing with scarce data, a common challenge in medical image analysis.

Future work should focus on expanding the dataset to include a wider variety of amyloidosis presentations and exploring advanced techniques to further reduce prototype bias. This will enhance the robustness and accuracy of FSL models in classifying renal amyloidosis and ensure better generalization across different manifestations of the disease.

Acknowledgements All authors belong to the Pathospotter project, which is partially sponsored by Fundação de Apoio à Pesquisa do Estado da Bahia (FAPESB), grant TO P0008/15 and Inova Fiocruz – Innovative ideas. Angelo Duarte is also sponsored by Universidade Estadual de Feira de Santana (UEFS), grant FINAPESQ TO 115/2024. Washington LC dos-Santos and Luciano Oliveira have research scholarships from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grants 306779/2017 and 308580/2021-4, respectively. This work was also partially supported by CAPES-PROAP 2023/2024 grants.

References

1. Abbas, Q.: An intelligent medical image classification system using few-shot learning. Concurrency and Computation: Practice and Experience **35**(2), e7451 (2023). https://doi.org/https://doi.org/10.1002/cpe.7451, https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.7451
2. Billion Polak, P., Prusa, J.D., Khoshgoftaar, T.M.: Low-shot learning and class imbalance: a survey. J. Big Data **11**(1) (Jan 2024)
3. Deng, Y., Li, H.: Deep learning for few-shot white blood cell image classification and feature learning. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(6),

2081–2091 (2023). https://doi.org/10.1080/21681163.2023.2219341, https://doi.org/10.1080/21681163.2023.2219341
4. Diallo, R., Edalo, C., Awe, O.O., A. Vance, E.: Machine learning evaluation of imbalanced health data: A comparative analysis of balanced accuracy, mcc, and f1 score. In: Awe, O.O., A. Vance, E. (eds.) Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA, pp. 283–312. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72215-8_12
5. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 594–611 (apr 2006). https://doi.org/10.1109/TPAMI.2006.79, https://doi.org/10.1109/TPAMI.2006.79
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015), https://api.semanticscholar.org/CorpusID:206594692
7. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. J. Big Data **6**(1) (Dec 2019)
8. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). pp. 730–734 (2015). https://doi.org/10.1109/ACPR.2015.7486599
9. Mahbub, T., Obeid, A., Javed, S., Dias, J., Hassan, T., Werghi, N.: Center-focused affinity loss for class imbalance histology image classification. IEEE Journal of Biomedical and Health Informatics **28**(2), 952–963 (2024). https://doi.org/10.1109/JBHI.2023.3336372
10. Medela, A., Picon, A., Saratxaga, C.L., Belar, O., Cabezón, V., Cicchi, R., Bilbao, R., Glover, B.: Few shot learning in histopathological images:reducing the need of labeled data on biological datasets. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 1860–1864 (2019). https://doi.org/10.1109/ISBI.2019.8759182
11. Megahed, F.M., Chen, Y.J., Megahed, A., Ong, Y., Altman, N., Krzywinski, M.: The class imbalance problem. Nature Methods **18**(11), 1270–1272 (Nov 2021)
12. Mienye, I.D., Sun, Y.: Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Informatics in Medicine Unlocked **25**, 100690 (2021). https://doi.org/https://doi.org/10.1016/j.imu.2021.100690, https://www.sciencedirect.com/science/article/pii/S235291482100174X
13. Ochal, M., Patacchiola, M., Vazquez, J., Storkey, A., Wang, S.: Few-shot learning with class imbalance. IEEE Transactions on Artificial Intelligence **4**(5), 1348–1358 (2023). https://doi.org/10.1109/TAI.2023.3298303
14. Raj, R., Mathew, J., Kannath, S.K., Rajan, J.: Crossover based technique for data augmentation. Computer Methods and Programs in Biomedicine **218**, 106716 (2022). https://doi.org/https://doi.org/10.1016/j.cmpb.2022.106716, https://www.sciencedirect.com/science/article/pii/S016926072200102X
15. Said, S.M., Sethi, S., Valeri, A.M., Leung, N., Cornell, L.D., Fidler, M.E., Herrera Hernandez, L., Vrana, J.A., Theis, J.D., Quint, P.S., Dogan, A., Nasr, S.H.: Renal amyloidosis: origin and clinicopathologic correlations of 474 recent cases. Clin. J. Am. Soc. Nephrol. **8**(9), 1515–1523 (Sep 2013)
16. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4080–4090. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

17. Sümer, Ö., Hellmann, F., Hustinx, A., Hsieh, T.C., André, E., Krawitz, P.: Few-shot meta-learning for recognizing facial phenotypes of genetic disorders. Stud. Health Technol. Inform. **302**, 932–936 (May 2023)
18. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
19. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/tan19a.html
20. Titoriya, A.K., Singh, M.P.: Few-shot learning on histopathology image classification. In: 2022 International Conference on Computational Science and Computational Intelligence (CSCI). pp. 251–256 (2022). https://doi.org/10.1109/CSCI58124.2022.00048
21. Tyagi, S., Mittal, S.: Sampling approaches for imbalanced data classification problem in machine learning. In: Singh, P.K., Kar, A.K., Singh, Y., Kolekar, M.H., Tanwar, S. (eds.) Proceedings of ICRIC 2019. pp. 209–221. Springer International Publishing, Cham (2020)
22. Walsh, R., Tardy, M.: A comparison of techniques for class imbalance in deep learning classification of breast cancer. Diagnostics **13**(1) (2023). https://doi.org/10.3390/diagnostics13010067, https://www.mdpi.com/2075-4418/13/1/67
23. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM Comput. Surv. **53**(3) (jun 2020). https://doi.org/10.1145/3386252, https://doi.org/10.1145/3386252
24. Yakupova, E., Bobyleva, L., Vikhlyantsev, I., Bobylev, A.: Congo red and amyloids: history and relationship. Bioscience Reports **39**(1) (2019). https://doi.org/https://doi.org/10.1042/BSR20181415, https://www.sciencedirect.com/science/article/pii/S1573493519002832
25. Zhang, R., Liu, Q.: Learning with few samples in deep learning for image classification, a mini-review. Front. Comput. Neurosci. **16**, 1075294 (2022)