AN2DL - Second Challenge Report
TheBigBatchTheory

Benedetta Mussini, Andrea Rossi, Fabio Rossi, Francesco Sarra

benedettamussini, Redsss, poliReus, fsarra

278186, 287278, 286400, 279755

January 25, 2026

1 Introduction

This report documents our journey in tackling the Second AN2DL Challenge: a fine-grained classification of histological slides into four subtypes `Luminal A`, `Luminal B`, `HER2(+)`, `Triple Negative`. Our work evolved from a simple baseline into a state-of-the-art pipeline, driven by a process of experimentation and debugging.

Initial attempts using standard Convolutional Neural Networks (CNNs) immediately highlighted the core challenge: severe overfitting. Our models excelled at memorizing the training data but failed to generalize, with validation scores finding a plateau at a low baseline. This discovery shifted our focus from simply training a model to a deeper investigation of our data, preprocessing, and architectural choices.

This report dives in that investigation. We detail our iterative approach, from data cleaning and advanced preprocessing techniques to the implementation of multi-scale and, finally, Multiple Instance Learning (MIL) architectures. Our final result reflects not just a single model, but an understanding of the problem's complexities.

2 Problem Analysis

The primary challenges of this project were two: data quality and task complexity:

- **Data Quality and Artifacts:** Our first priority was data cleaning. We noticed that many slides were contaminated with stains (*"slime"*) or other artifacts (*Shrek* images). We invested considerable effort in developing an automated cleaning pipeline using *HSV color-space filtering* to remove artifacts and correcting stained slides to recover them for use. This was a critical learning step, even though we later found that many stained images were duplicates and could be removed.

- **The Overfitting Problem:** Our early models, even on clean data, showed severe overfitting. The validation F1 score would not improve beyond a low baseline ($\sim$0.3), while the training score quickly reached perfection. This proved that the classification task was highly challenging. We hypothesized that the visual signal in a single, small image patch was too weak and ambiguous for a model to distinguish between the subtle differences of the four subtypes. This realization forced us to explore more advanced methods to provide the model with better context.

3 Method

Our approach evolved through three main stages, each designed to solve the problems identified in the previous one.

Stage 1: Baseline CNN with Sliding-Window Tiling In our initial approach we used a sliding window to extract 512x512 pixel tiles from the entire tissue area. These tiles were fed to a pre-trained `ResNet-18` model. This method immediately presented two challenges: class imbalance and high variance in tile quality.

To address the imbalance between the four classes, we implemented class weighting in the loss function (`CrossEntropyLoss`), assigning a higher weight to minority classes. Furthermore, to prioritize tiles with more diagnostic potential, we experimented with a `WeightedRandomSampler`. This sampler gave a higher probability of being selected to tiles containing a larger amount of tissue.

Despite these efforts, this approach ultimately failed due to massive overfitting. We concluded that the naive tiling created too much ”label noise” (e.g., background tiles with positive labels) and that these sampling strategies were insufficient to overcome the poor quality of the data generated. This led us to abandon the sliding-window method entirely.

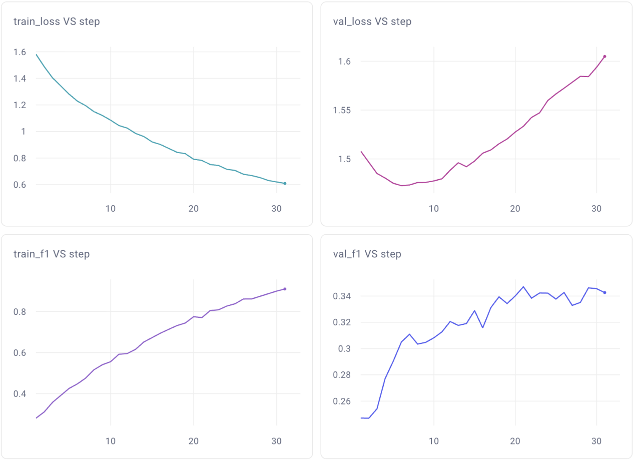


Figure 1: Stage 1 - Severe overfitting with the baseline CNN, despite using class weights and sampling strategies

Stage 2: ROI-Centric and Multi-Scale Tiling
To improve the signal quality, we developed a ROI-centric preprocessing script. Instead of tiling the whole slide, we used the provided masks with `cv2.findContours()` to extract tiles centered only on the diagnostically relevant regions.

This improved the results, but overfitting was still a major issue. We then implemented a `Dual-Branch Multi-Scale` model. This model processed two tiles for each ROI simultaneously:

- A 768x768 tile to capture architectural context.
- A 256x256 tile to capture cellular detail.

This strategy led to our first significant breakthrough, pushing the validation F1 score above 0.40 by providing the model with a much richer input. However, the result on the test set was still disappointing, meaning the model was not actually generalizing.

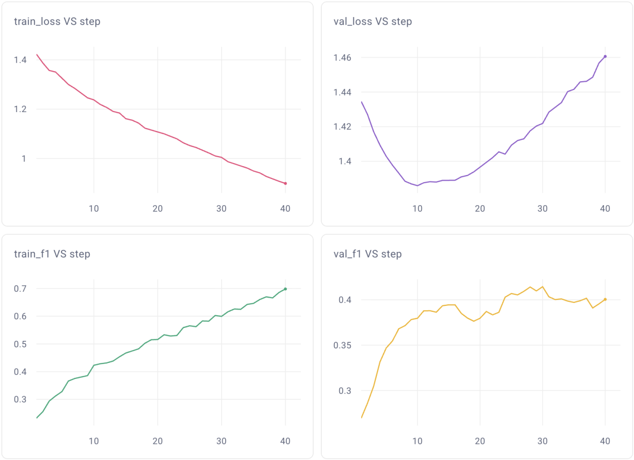


Figure 2: Stage 2

Stage 3: Multiple Instance Learning (MIL)
While multi-scale improved tile-level prediction, it still relied on averaging predictions. To make a true slide-level diagnosis, we implemented our final architecture: an Attention-based MIL model. This model treats all tiles from a slide as a single ”bag”. The core components are:

- **Backbone:** A Vision Transformer (`vit_small_patch16`) acts as a feature extractor for each tile.
- **Aggregation:** An attention mechanism learns to weigh the most important tiles in the bag, creating a single, powerful feature vector for the entire slide.

This approach makes a single, holistic prediction per slide and represents the state-of-the-art for this problem.

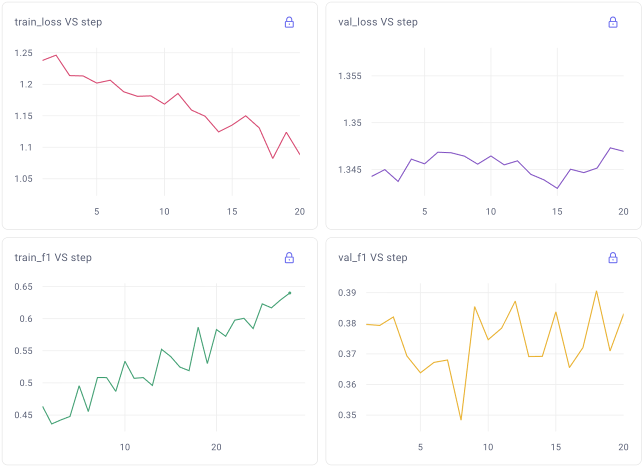


Figure 3: Stage 3, here the learning curves of our best model are shown

4 Results

We conducted a series of experiments to guide our architectural decisions. Each new approach was compared to the previous ones.

Table 1: Model performance across architectures.

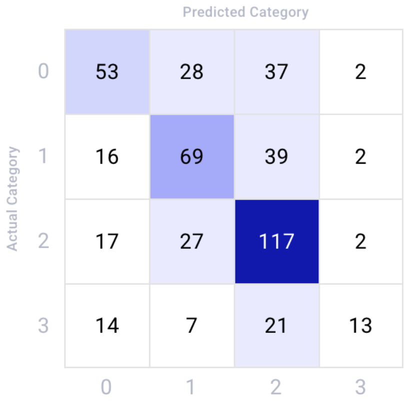
Model	Val F1	Test F1
ResNet-18 (SW)	0.35	0.27
ResNet-18 (ROI+MS)	0.424	0.340
AttnMIL (ViT-S)	**0.3906**	**0.3898**

The results in Table 1 clearly demonstrate our progression. The initial baseline suffered from a large gap between validation and test, indicating poor generalization. The introduction of ROI-centric, multi-scale data improved the validation score and reduced the gap. Finally, the Attention-MIL architecture provided the best and most robust performance on the test set, confirming its superiority for this slide-level task.

Throughout all stages, we systematically tuned hyperparameters. The most effective training strategy involved a two-phase fine-tuning approach combined with a low learning rate (5e-6) and a `CosineAnnealingLR` scheduler, which was crucial for stabilizing the training of these complex models. Figure 3 shows the learning curves from our best experiment, highlighting a stable learning process

5 Discussion

Our final F1 score of 0.3898, while modest, represents a significant achievement built upon a foundation of systematic debugging. The journey from a failing baseline to a stable MIL pipeline demonstrates that the primary challenge was not a lack of signal, but the difficulty in extracting it. The confusion matrix of our best model still shows significant overlap between some of the classes, suggesting that even a powerful slide-level approach struggles with the inherent visual ambiguity of the task and the relation between those classes could be investigated more deeply. The biggest weakness of our final model is likely still the information bottleneck: it learns from pre-selected ROIs, but might miss crucial information in the parts of the tissue that were not masked.



	0	1	2	3
0	53	28	37	2
1	16	69	39	2
2	17	27	117	2
3	14	7	21	13

Figure 4: Confusion matrix of our best model.

6 Conclusions

In this project, we successfully navigated the challenges of a fine-grained histological classification task. We demonstrated the failure of simple approaches and iteratively developed a state-of-the-art Attention-MIL pipeline. Our key contribution is the documentation of this methodical process, highlighting the importance of ROI-centric preprocessing and slide-level aggregation. Our final result underscores the profound difficulty of the problem. Future work should explore the exploiting of more advanced models as pretrained backbones. Also, the usage the entire slide for tiling (not just the ROIs) within the MIL framework could allow the attention mechanism itself to discover the most diagnostically relevant regions.