Development and Implementation of Virtual Staining Techniques in Bone Marrow Samples using Deep Learning

I. Hernández-Abad[1,2], J.E. Ortuño[2,1], A. Mendoza[3], M. Gómez-Álvarez[4], A. Ortiz-Ruiz[5], A.Basterra-García[1,2], D. Bermejo-Peláez[6], D. Brau-Queralt[6] N. Díez[6] , A. Santos[1,2], C. Benavente[4], J.Martínez-López[3], M.Linares[5], M. Luengo-Oroz[6], M.J. Ledesma-Carbayo[1,2]

[1] Biomedical Image Technologies, Universidad Politécnica de Madrid, Madrid, Spain {ignacio.hernandez.abad, mj.ledesma}@upm.es

[2] CIBER de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Madrid, Spain

[3] Servicio de Hematología y Hemoterapia, Hospital Universitario 12 de Octubre, Madrid, Spain

[4] Servicio de Hematología y Hemoterapia, Hospital Clínico San Carlos, Madrid, Spain

[5] Universidad Complutense de Madrid, Madrid, Spain

[6] Spotlab, Madrid, Spain

Summary

This work aims to enhance diagnostic processes in hematology through virtual staining techniques. This approach seeks to overcome the limitations of traditional staining methods, which often demand multiple bone marrow aspirations and can be time-consuming leading to delays in diagnostic and specialized treatment. The objective is to develop a virtual staining tool that can transform morphological stains like May-Grünwald Giemsa into histochemical stains such as Perls´ stain. Our methodology employs unsupervised cycle-consistent generative adversarial networks (CycleGAN) to create virtual stains from bone marrow smear images. This architecture is based on generators and discriminator networks that are trained simultaneously in a competitive manner, each striving to outperform the other. This work includes protocols for acquiring high-resolution images of bone marrow samples, the specifications of a ring sideroblast-conditioned dataset class created to modify the pairing of images during the training process, comparison of results from various implemented architectures, and detailed procedures for obtaining and evaluating results. The obtained results show a promising capacity to generate realistic virtual images that can successfully pass a Visual Turing Test, convincing experts that they are genuine. Furthermore, the findings reveal which of the seven tested configurations of the proposed model is more effective in preserving pathological information of the images. This work presents a promising advancement in medical diagnostics, leveraging deep learning to streamline and enhance the virtual staining process for bone marrow samples.

1. Introduction

Myelodysplastic Syndromes (MDS) are diseases marked by ineffective hematopoiesis, cytopenia or morphological dysplasia in the bone marrow. These issues impair oxygen distribution, infection defense, and blood clot formation. The main clinical concerns with MDS are the morbidities from cytopenia and its high risk of progressing to Acute Myeloid Leukemia (AML), a rapidly advancing blood cancer and the most common acute leukemia in adults [1]

A key diagnostic test for MDS is a complete blood count on bone marrow smears to assess the ratio of undifferentiated cells. When Ring Sideroblasts (RS) account for ≥15% of the erythroid precursors, myelodysplastic syndrome with RS (MDS-RS) is diagnosed. To visualize and identify the cells under a microscope, a morphological stain such as May-Grünwald Giemsa (MGG) is employed [2].

Additionally, histochemical stains can identify cellular abnormalities aiding in disease classification. For instance, Perls' stain dyes precipitated iron granules in the erythroblasts, helping identify Ring Sideroblasts (RS). Sideroblasts are classified as type 0,1 and 2 (denoted as negative RS in this work) or type 3 (positive RS) based on granule count and distribution [3]. Positive RS have at least 5 iron granules in a perinuclear distribution as shown in *Figure 1*.

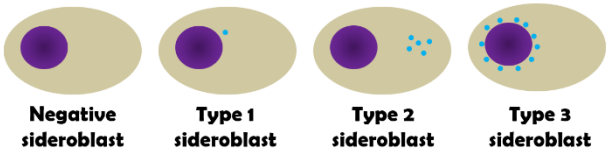

Figure 1. *Schematic of sideroblast classification.*

This work hypothesizes that the relationship of morphological traits and RS signs could be learned by a deep learning model. In this work we implement a generative framework to virtually generate the corresponding Perl's staining, thereby facilitating hematologists in making more accurate diagnoses and prognoses of MDS.

2. Materials & Methods

Images were acquired using Spotlab's digitalization system, which includes a smartphone attached to a 3D-printed microscope adapter and linked to a mobile app. The Hematology and Hemotherapy Service of Hospital 12 de Octubre in Madrid digitized samples from 15 patients capturing 40 MGG and 20 Perls fields of view per patient. The dataset was divided into 12 patients for training and 3 for testing. A total of 1842 and 1418 patches containing erythroblasts, in MGG and Perls respectively, with a patch size of 256×256 were extracted. A mean of 124±63 (MGG) and 95±96 (Perls) patches were obtained per patient.

ISBN: 978-84-09-67332-2

Hematologists labeled the extracted patches as negative or positive RS according to the criteria followed in *Figure 1*. The dataset distribution is detailed in *Table 1*.

Stain	Split	#Negative patches	#Positive patches	# Total patches
MGG	Train	1527	110	1637
	Test	140	65	205
Perls	Train	791	499	1290
	Test	63	65	128

Table 1. *Patch Dataset Distribution.*

There has been implemented two kinds of datasets, the original implementation of the dataset class which randomly pairs images from each domain, and our proposed approach based on pairing images conditioned by their labels. We introduce a new class, termed the RS-Conditioned dataset class, which ensures that images from both stains are only paired with those having the same label in the other stain. This idea is illustrated in *Figure 2*.

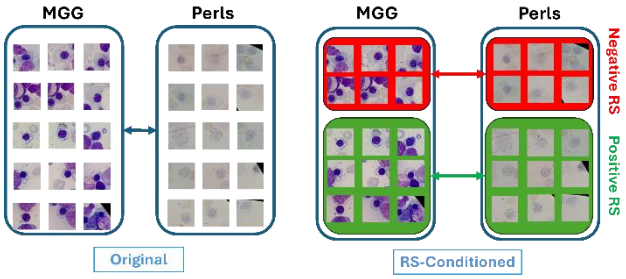

Figure 2. *Original vs. RS-Conditioned dataset classes.*

An unsupervised learning architecture, Cycle-Consistent Generative Adversarial Networks (CycleGAN) [4], was adapted and tested. This model uses two generators for stain transfers between domains (A→B and B→A) and two discriminators to distinguish real from fake images. CycleGAN employs a cycle loss to minimize the difference between the original and the reconstructed image.

Seven experiments were conducted to determine the best configuration for virtual staining. The experiments varied the dataset type (Original vs. RS-Conditioned), generator type (ResNet vs. U-Net) and the number of discriminator layers. All models were trained for 200 epochs (100 epochs with fixed learning rate (l_r) of 0.0002 and 100 with a linear l_r decay). The ResNet architecture is composed of 9 residual blocks and the U-Net has 8 downsamplings.

Experiment #		1	2	3	4	5	6	7
Dataset Class	Original	✓						
	RS-Conditioned		✓	✓	✓	✓	✓	✓
Generator	ResNet	✓	✓		✓	✓		
	U-Net			✓			✓	✓
Discriminator layers		3	3	3	2	5	2	5

Table 2. *Experiment Configurations*

A Visual Turing Test (VTT) [5] was developed to evaluate the performance of each experiment. Each VTT consists of a compilation of 100 randomly selected images, equally divided between real and generated images, and between positive and negative RS. For each image, two responses are required: whether the image is believed to be real or fake, and whether the cell corresponds to a positive or negative RS.

Figure 3. *Example of VTT image evaluation presented to the hematologists.*

Two hematologists from different hospitals performed the seven VTTs and employing their answers and the corresponding labels of the patches two confusion matrices were created each time. These confusion matrices allow to visualize the performance of the models in two tasks: the realism of the images and the transfer of the pathological information.

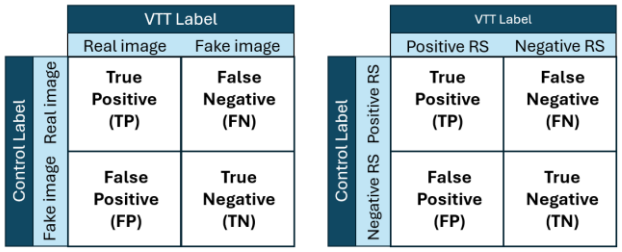

Figure 4. *Confusion matrices for evaluation tasks.*

Regarding the evaluation of the image realism quality, the metric of interest is the False Positive Rate (FPR) which indicates how many fake images tricked the experts and were classified as real. On the other hand, for the evaluation of the pathological information transference, the metric of interest is the accuracy, as the objective is that the experts correctly classify the sideroblasts from the Perls virtually generated image.

Additionally, the Frechét Inception Distance (FID) [6] was calculated as an agnostic metric to assess the distance between pairs of real vs. generated datasets. For the shake of comparison this metric was computed in different datasest: train, test, train + tests and the subcohort of train + test employed to generate the VTTs. For the fake images, the corresponding datasets generated for each experiment were used. Subsequently, Pearson´s correlation coefficient was estimated to determine if there is a correlation between the FID and the FPR calculated in the VTTs. Ideally, a lower FID indicates more realistic generated images , and consequently, a correlation coefficient of -1 would be expected.

ISBN: 978-84-09-67332-2

3. Results & Discussion

Inferences were made to obtain virtually stained Perls from real MGG for each experiment. Some examples of these inferences are shown in the *Figure 5*. Later, a random selection from the pool of both generated and real Perls were selected to create the VTTs performed by the hematologists.

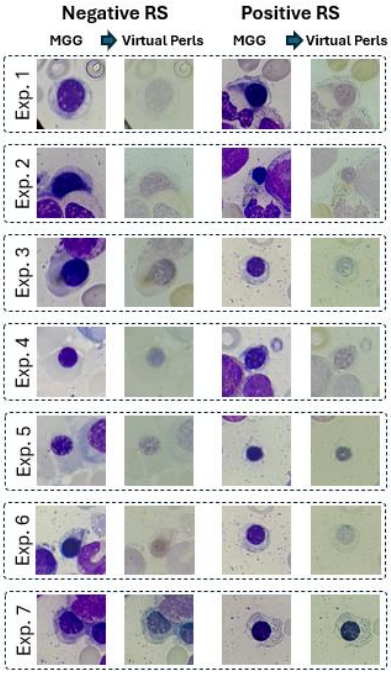

Figure 5. *Inference examples of MGG to virtual Perls*

Regarding the first task involving the image realism, *Figure 6* shows the values for FPR and FNR. The FPR indicates how well the model was able to generate realistic images that could deceive experts into believing they are real. It is observed that the second experiment, which employs the RS-conditioned dataset class performs the best in this task, with 68% of fake images classified as real. Additionally, this same experiment shows the highest FNR that would indicate that the experts tried to compensate the number of images classified as real and ended labeling actual real images as fake. Additionally, in the experiments where a 5-layers discriminator is employed (experiments 5 & 7) the generator is not able to beat the discriminator and in consequence, lower FPR are obtained.

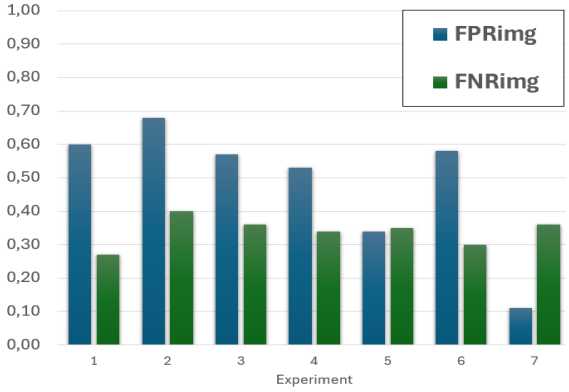

Figure 6. *FPR (blue) & FNR (green) values per experiment.*

If we observed the *Table 3*, it can be appreciated that values of FID in the test dataset are always significantly worse than in the training dataset, this would indicate an overfitting. Additionally, it must be clarified that 1 of the 3 test cases has a defective stain that could also affect this metric. It is also highlighted in *Table 3* how experiment 1 beats the rest and it would indicate the relevance of more possible combinations of paired images when the dataset class is not conditioned. FID and FPR of the VTTs have a Pearson's correlation coefficient of -0.74 which reflects that this agnostic metric correctly reflects the experts' opinions.

Experiment	Test	Train	Train + Test	VTT subcohort
1	**111.676**	**72.232**	**72.305**	79.054
2	115.044	72.395	72.533	**78.985**
3	118.972	78.178	78.276	85.216
4	118.127	76.078	76.123	83.831
5	202.733	145.268	147.304	154.509
6	116.841	77.301	77.205	84.746
7	159.656	110.639	111.878	118.573

Table 3.*FID results. Best experiment for each dataset in* ***bold****. Best dataset for each experiment underlined.*

Moving to the pathological information transfer task, it is observed in the *Figure 7* that when calculating the accuracy of RS positive detection in both, real and fake images, this value always oscillate between 0.6-0.7. When analyzing these images separately, it is appreciated that the mean accuracy value for real images is 0.77 which should be the target to aspire in the fake ones. Unfortunately, this value falls to 0.5 when employing fake images.

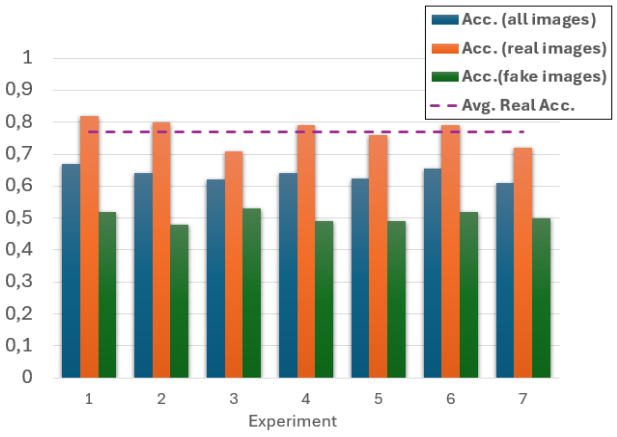

Figure 7. *Accuracy values employing all images (blue), only real (orange) and only fake images (green).*

Taking a further look at why these values in fake images are so low, it can be appreciated when calculating the corresponding FNR and FPR error rates in the *Figure 8* that the models tend always to fail in the inference of images compatible with positive RS that could be justified due to the unbalanced training dataset. Nevertheless, it is observed that having more possible combinations (original dataset) reduces the value of FNR with respect to employing the conditioned dataset. On the other hand, it is also seen that the RS-Conditioned pairing reduces the

ISBN: 978-84-09-67332-2

hallucinations that virtually stain negative RS as positives, and in consequence lower values of FPR are obtained.

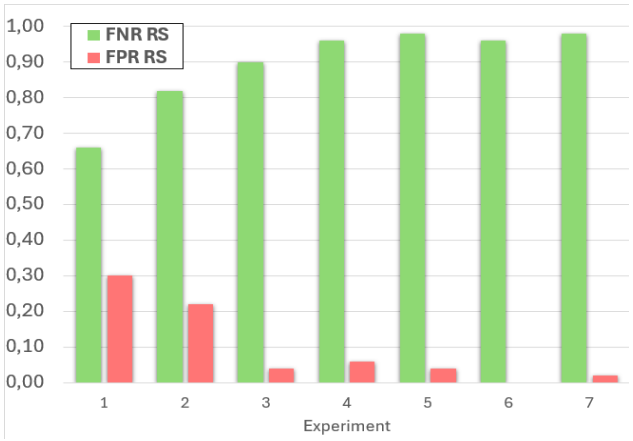

Figure 8. *FNR and FPR error rates of fake images per experiment.*

A final observation has been performed when calculating the Root Mean Squared Error (RMSE) between the estimated and the actual percentages of positive RS per case. The results indicate that utilizing an RS-conditioned dataset class is advantageous for inference in cases with lower percentages of positive RS. This approach reduces the occurrence of hallucinations in negative RS, which are underrepresented in the dataset. In contrast, cases with higher percentages of positive RS derive less benefit from the conditioned pairing, as they are less affected by the misclassification of negative patches as positive.

4. Conclusions

In this work, a virtual staining method has been developed to synthesize Perls images from MGG which are capable to deceive experts and that are promising for assisting in the diagnosis MDS.

The results indicate that, of the seven adaptations implemented of the CycleGAN architecture, employing a ResNet generators and 3-layers discriminators is the most effective for our needs.

Additionally, the proposed approach employing a RS-Conditioned dataset class shows clear improvements. The model tends to not only reduce hallucinations that cause to classify negative RS as positive (FP) but also improve the realism quality of the images overall.

On the other hand, having more image pair combinations, as seen in the original dataset, helps reduce false negatives (positive RS classified as negative). This difference is expected to decrease with a larger, more balanced RS-Conditioned dataset that includes more positive patches and diverse data augmentations.

This work also identified a significant need for an agnostic metric to reduce the dependency of hematologist supervision performing VTTs, which are tedious, time-consuming and imply a huge variability due to the subjectivity of the experts who perform them.

Finally, it is noteworthy to highlight the beneficial effects that would bring the employment of virtual staining in the clinical routine. This methodology would be a time and cost-effective solution in comparison to chemical stains that would speed up the diagnosis of patients significantly and reduce the workload of hospitals' technicians. Additionally, digital staining would also mean a reduction of toxic chemical waste and the exposure of the technicians to these agents. Finally, it would require less bone marrow aspirate material and could provide extra diagnostic information to the hematologists in cases where these smears would not be available otherwise.

Acknowledgments

This project has received partial funding from the European Union's Horizon 2020 research and innovation program (grant agreement No. 881062); the European Union - NextGenerationEU through the Spanish Government's "Plan de Recuperación, Transformación y Resiliencia"; the Instituto de Salud Carlos III (ISCIII) under grants PMPTA22/00169, PMPTA22/00088, PMPTA22/00041, PMPTA22/00023, PMPTA22/00101; and the Centro para el Desarrollo Tecnológico y la Innovación (CDTI) under grant EXP 00156466 / IDI-20230066.

References

[1] L. Adès, R. Itzykson, and P. Fenaux, "Myelodysplastic syndromes," *The Lancet*, vol. 383, no. 9936, pp. 2239–2252, Jun. 2014, doi: 10.1016/S0140-6736(13)61901-7.

[2] R. P. Hasserjian, "Acute myeloid leukemia: Advances in diagnosis and classification," *Int J Lab Hematol*, vol. 35, no. 3, pp. 358–366, Jun. 2013, doi: 10.1111/IJLH.12081.

[3] G. J. Mufti *et al.*, "Diagnosis and classification of myelodysplastic syndrome: International Working Group on Morphology of myelodysplastic syndrome (IWGM-MDS) consensus proposals for the definition and enumeration of myeloblasts and ring sideroblasts," *Haematologica*, vol. 93, no. 11, pp. 1712–1717, Nov. 2008, doi: 10.3324/HAEMATOL.13405.

[4] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 2242–2251, Mar. 2017, doi: 10.1109/ICCV.2017.244.

[5] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual Turing test for computer vision systems," *Proc Natl Acad Sci U S A*, vol. 112, no. 12, pp. 3618–3623, Mar. 2015, doi: 10.1073/pnas.1422953112.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 6627–6638, Jun. 2017, doi: 10.18034/ajase.v8i1.9.