Bone Marrow Fibrosis Grading Using Attention-Based Multiple Instance Learning

Lauren M. Zuromski[1] LAUREN.ZUROMSKI@ARUPLAB.COM
[1] *Institute for Research and Innovation, ARUP Laboratories, Salt Lake City, UT*
Alexandra E. Rangel[1] ALEXANDRA.RANGEL@ARUPLAB.COM
Muir J. Morrison[1] MUIR.MORRISON@ARUPLAB.COM
Paul M. English[1] PAUL.ENGLISH@ARUPLAB.COM
Nicholas C. Spies[1,2] NICK.SPIES@ARUPLAB.COM
Brendan O'Fallon[1] BRENDAN.O'FALLON@ARUPLAB.COM
[2] *Department of Pathology, University of Utah, Salt Lake City, UT*
David P. Ng[1,2] DAVID.NG@ARUPLAB.COM

Editors: Accepted for publication at MIDL 2025

Abstract

An attention-based multiple instance learning approach is used to improve bone marrow fibrosis (BMF) grading from whole slide images of bone marrow core biopsies. Slide-level labels were parsed from biopsy reports using a large language model, and features were extracted using our fine-tuned DINOv2-based backbone. The model achieved good agreement between BMF predictions and labels ($R^2 = 0.72$, $\kappa = 0.58$). Attention maps showed the model focused on diagnostic regions, highlighting its accuracy and interpretability.

Keywords: Bone marrow fibrosis, multiple instance learning, whole slide imaging, digital pathology, large language model, DINOv2

1. Introduction

Bone marrow fibrosis (BMF) is a condition where reticulin fibers accumulate, thickening the marrow and impairing blood cell production. BMF is a hallmark of myelofibrosis (MF), a rare myeloproliferative neoplasm (MPN) with prognosis influenced by fibrosis grade, age, genetic mutations, and risk of progression to leukemia. Accurate grading is essential for assessing severity and guiding treatment. BMF is commonly graded on an integer scale from 0 to 3, ranging from no to extensive reticulin fiber intersections (Thiele et al., 2005). Grading relies on qualitative assessments by pathologists examining bone marrow (BM) core biopsy whole slide images (WSIs), assigning the highest grade present in $\geq 30\%$ of the marrow. Although there is general agreement among experts (Kvasnicka et al., 2016), the BMF grading scale poorly handles fibrosis heterogeneity (Ryou et al., 2022).

To address this, a recent study estimated BMF on a continuous scale by modeling ranks of fibrosis severity, improving MPN subtype differentiation and risk stratification (Ryou et al., 2022). However, this method required expert review of thousands of tiles pairs, and the slide-level BMF score was simply calculated as the average of the tile-level grades.

In our ongoing study, we refine fibrosis grading through an attention-based multiple instance learning (ABMIL) approach, where the model learns image tiles that are important

using only slide-level labels. Features were derived using our fine-tuned backbone, and slide-level labels were extracted from biopsy reports using a large language model (LLM). Our model highlights diagnostically relevant regions and shows strong performance, suggesting improved objectivity and predictive power over prior methods.

2. Methods

For this analysis, we used clinical WSIs and core biopsy reports spanning 2019–2024 from our reference laboratory. The WSIs were digitized from slides, and the biopsy reports were retrieved from internal databases.

BMF grades were extracted from biopsy reports using Ollama, an open-source implementation of the Llama 3.1 70B LLM (Touvron et al., 2023; Grattafiori et al., 2024). Fibrosis grades were extracted from report text, with heterogeneous descriptions (e.g., "MF-2 to focally MF-3") assigned intermediate values (e.g., 2.5) to align with discrete labels.

WSIs of reticulin-stained BM core biopsies for each BMF-labeled case were tiled to generate the evaluation dataset. An object detection model isolated core tissue to guide recursive tiling from low to high magnification, removing background via color filtering. The evaluation dataset included 322 training and 80 testing cases, each using 512 tiles (224×224 pixels at 40x optical objective) that are down- or upsampled as needed.

We used an ABMIL framework, allowing the model to learn from image tile sets with only slide-level labels (Ilse et al., 2018). Features were extracted using a ViT-L/14 backbone fine-tuned with a DINOv2-based approach (Oquab et al., 2023), trained on millions of core tiles from various stains and magnifications, and passed to the ABMIL framework. DINOv2 leverages self-supervised learning through vision transformers, capturing pathology-specific features without labels. For benchmarking, we used common histopathology backbones: Virchow2 (Zimmermann et al., 2024), UNI (Chen et al., 2024), and Prov-GigaPath (Xu et al., 2024). Model optimization used a batch size of 16, learning rate of 0.00025, and early stopping based on lowest test-set RMSE.

3. Results and Discussion

The ABMIL model using our in-house backbone performed comparably to the benchmarks, achieving an RMSE of 0.52 (Table 1). Since the in-house backbone is intended for production use, it was necessary to demonstrate performance comparable to widely used histopathology backbones. Ongoing efforts of scanning more cases to expand the training sets for both the ABMIL model and backbone could help improve model performance.

We evaluated the ABMIL model trained with our in-house backbone on both continuous and discrete fibrosis grade scales. There is a strong correlation between predicted and actual values (R^2=0.72, $p < 0.0001$, Figure 1). For categorical analysis, we excluded intermediate labels and rounded predictions, yielding 67 cases. Predictions deviate from actual grades by one grade at most (Figure 1). Cohen's Kappa was 0.58 [0.42–0.73], indicating moderate agreement between predictions and labels, which exceeds expert inter-rater variability (κ=0.51 [0.48–0.54]) (Kvasnicka et al., 2016).

ABMIL models offer interpretability via attention maps, highlighting regions influencing predictions. In a correctly predicted grade-three case, the model correctly focuses on regions

with extensive fibrosis while down-weighting bone and adjacent tissue, which are areas known to skew severity estimates (Figure 2).

Table 1: Model performance (RMSE [95% CI]) across different backbones.

Backbone	In-house	UNI	Virchow2	Prov-GigaPath
RMSE [95% CI]	0.52 [0.45–0.61]	0.53 [0.46–0.63]	0.54 [0.47–0.64]	0.54 [0.47–0.65]

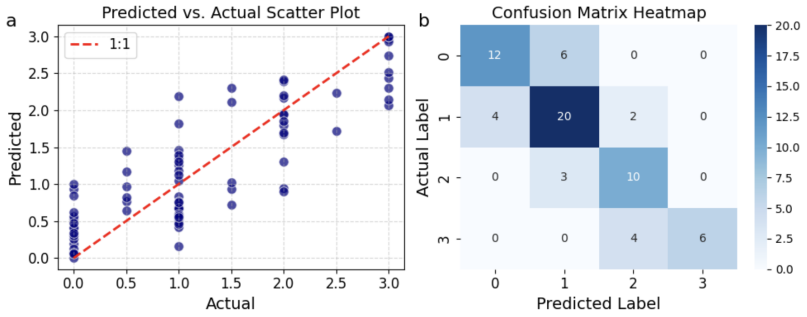

Figure 1: a) Predicted vs. actual labels using the in-house backbone (R^2=0.72, $p < 0.0001$). b) Confusion matrix showing agreement between actual and rounded predictions.

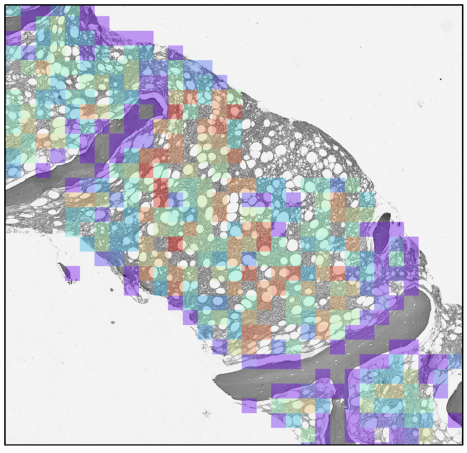

Figure 2: BM core with overlaid high/low model attention shown in warm/cool colors.

Acknowledgments

We thank ARUP Laboratories for enabling this study, especially the R&I Digital Imaging Center for data and slide digitization support.

References

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30: 850–862, 2024. doi: 10.1038/s41591-024-02857-3. URL https://doi.org/10.1038/s41591-024-02857-3. Published 19 March 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan,

Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli,

Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. Eprint arXiv:1802.04712, 2018.

Hans-Michael Kvasnicka, Christine Beham-Schmid, Robert Bob, Stephan Dirnhofer, Kamal Hussein, Hans Kreipe, et al. Problems and pitfalls in grading of bone marrow fibrosis, collagen deposition and osteosclerosis – a consensus-based study. *Histopathology*, 68(6): 905–915, 2016. doi: 10.1111/his.12871.

Maxime Oquab, Théo Darcet, Thibaut Moutakanni, Seyed-Mohsen Dezfooli, Daniel Haziza, Javier Suárez, Julien Mairal, Piotr Bojanowski, Ivan Laptev, Natalia Neverova, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Hosuk Ryou, Korsuk Sirinukunwattana, Alan Aberdeen, Gillian Grindstaff, Bernadette J Stolz, Helen Byrne, Heather A Harrington, Nikolaos Sousos, Anna L Godfrey, Claire N Harrison, Bethan Psaila, Adam J Mead, Gabrielle Rees, Gareth D. H. Turner, Jens Rittscher, and Daniel Royston. Continuous indexing of fibrosis (cif): improving the assessment and classification of mpn patients. *Leukemia*, 37(2):348–358, 2022. doi: 10.1038/s41375-022-01773-0. Published 2022 Dec 5. Correction: Leukemia. 2023 Jan 12;37(2):503. PMID: 36470992, PMCID: PMC9898027.

Jürgen Thiele, Hans Michael Kvasnicka, Fabio Facchetti, Vito Franco, Jon van der Walt, and Attilio Orazi. European consensus on grading bone marrow fibrosis and assessment

of cellularity. *Haematologica*, 90(8):1128–1132, 2005. URL https://pubmed.ncbi.nlm.nih.gov/16079113/. PMID: 16079113.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w. URL https://www.nature.com/articles/s41586-024-07441-w.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024. URL https://arxiv.org/abs/2408.00738.