German Medical Data Sciences 2025: GMDS Illuminates Health
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI251413

Classifying the AMi-Br Mitotic Figure Dataset with AUCMEDI

Daniel HIEBER[a,b,c], Friederike LISCHE-STARNECKER[a], Johannes SCHOBEL[b], Rüdiger PRYSS[c], Frank KRAMER[d], and Dominik MÜLLER[d,e,1]
[a] *Department of Neuropathology, Pathology, Medical Faculty, University of Augsburg*
[b] *DigiHealth Institute, Neu-Ulm University of Applied Sciences*
[c] *Institute of Medical Data Science, University Hospital of Würzburg*
[d] *Faculty of Applied Computer Science, University of Augsburg*
[e] *Institute for Digital Medicine, University Hospital Augsburg*
ORCiD: DH 0000-0002-6278-8759, FLS 0000-0003-1948-1580, JS 0000-0002-6874-9478, RP 0000-0003-1522-785X, FK 0000-0002-2857-7122, DM 0000-0003-0838-9885

Abstract. Introduction: Mitotic figure (MF) density has been established as a key biomarker for certain tumors. Recently, the differentiation between atypical MFs (AMF) and normal MFs (NMFs) has gained increased interest in research, as AMFs density could be an independent biomarker. This results in the challenge of finding an automated, deterministic way to differentiate between AMFs and NMFs. **Methods:** In this study, the AUCMEDI deep learning framework is applied to the recently published AMi-Br dataset to get a first bearing on the complexity of the task at hand. The dataset includes eight mitotic subclasses derived from breast cancer samples, four for NMFs and four for AMF. Using a patient-level cross-validation strategy and a ConvNeXt-based ensemble, we trained and evaluated an eight-class subtype classification model. **Results:** Our results show high specificity across all classes (≥ 90%), but sensitivity varies significantly between mitotic subclasses (0–82%), reflecting the dataset's inherent challenges. The mean AUC of 85.90% outperforms the binary classification baseline (69.8%). **Conclusion:** The results highlight the promise of progress in subclass-level mitotic analysis while pointing to areas for further model refinement.

Keywords. Atypical Mitotic Figures, Deep Learning, Classification, Computational Pathology, Computer Vision

1. Introduction

In the histopathological assessment of some tumors, mitotic figure (MF) quantification plays a central role in evaluating tumor proliferation and predicting patient outcomes [1,2]. As with many aspects in pathology, this process traditionally relies on expert visual inspection of hematoxylin and eosin-stained slides, which is time-consuming and subject to inter-observer variability [3]. In recent years, automated image analysis tools powered by deep learning have demonstrated promising results in standard MF detection and classification [4,5].

[1] Corresponding Author: Dominik Müller, University of Augsburg, Universitätsstraße 2, 86159 Augsburg, Germany, E-Mail: dominik.mueller@informatik.uni-augsburg.de.

While most studies focus on the identification of mitotic figures in pathological images, recent evidence suggests that atypical mitotic figures (AMFs) may serve as a valuable prognostic marker [6]. However, distinguishing AMFs from normal mitoses (NMFs) is a significantly more complex task due to overlapping morphological features and the rarity of atypical cases.

To facilitate research on AMF classification, the AMi-Br dataset was recently introduced [7] offering a foundation for developing and evaluating novel approaches, as well as providing a baseline for comparison and validation. It combines mitotic figure samples from the revised TUPAC [8] and MIDOG 2021 [9] and 2022 [10] datasets. It is the first dataset that includes a fine-grained eight-class label scheme distinguishing both normal and atypical subtypes. In this work, we [11]developed a deep learning based model for performing subtype MF classification on the AMi-Br dataset. Our primary objective is to assess the model's performance across these fine-grained classes and provide baseline insights on the complexity of the challenge for future work in this domain.

2. Material and Methods

The study is based on the AMi-Br dataset containing 3,720 annotated image patches of MF [7]. The patches are split into 832 AMFs and 2,888 cases of NMFs. Each type of MFs has four subtypes additionally labeled, namely for typical MFs: prometaphase (1,527 tiles), metaphase (1,349 tiles), ring shape (52 tiles), and ana- & telophase (253 tiles). AMFs are separated into bipolar asymmetry (59 tiles), tri-/multipolar asymmetry (71 tiles), segregation abnormalities (154 tiles), and a fourth group "other" (255 tiles) for all remanning cases. This data setup allows multiple possible ML-tasks on the dataset. In this work the focus is on the subtype classification task, therefore, a total of eight possible classes are to be considered. Figure 1 shows example images for each of the 8 different classes.

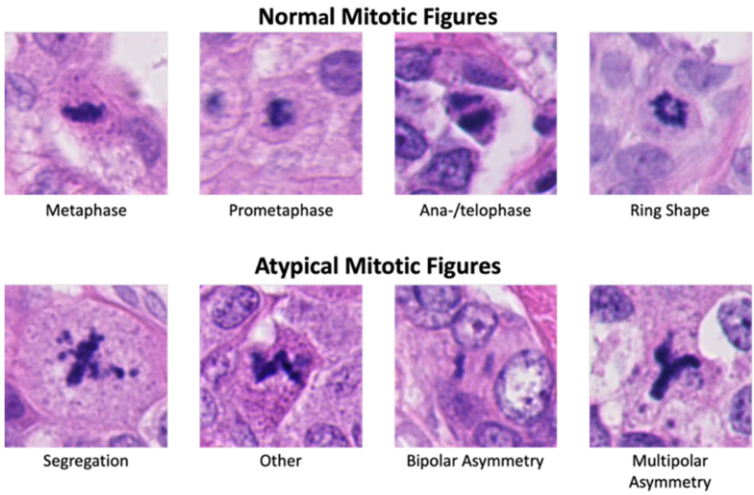

Figure 1. Overview of the eight different Classes in the AMi-Br dataset.

The dataset is split into training and testing data using an 80/20 patient-wise split. With this approach patients are either in the test or training set, but images from the same patient cannot be in both. While this results in worse results as showcased by the original work [7], the resilience of the model is increased and the risk of overfitting decreased. As preprocessing, image resolutions were scaled to 224×224 pixel and intensity values were zero-centered using Z-score normalization, with the mean and standard deviation derived from the ImageNet dataset [12].The subtype-classification of the data was performed using the AUCMEDI framework [11], allowing the creation of a deep learning based streamlined pipeline . Based on a ConvNeXt model architecture [13], a three-fold cross-validation approach was used to train a model ensemble for the classification task. The batch size was defined as 42 with a fixed amount of 180 batches per epoch. A transfer learning strategy was employed, beginning with training only the head layers of the network for the first 10 epochs using a fixed learning rate of $1e^{-4}$. After this initial phase, the entire network was fine-tuned using a dynamic learning rate that gradually decreased from $1e^{-4}$ to $1e^{-7}$ over a maximum of 1000 epochs. To avoid overfitting, early stopping was used, halting training if the validation loss did not improve for 10 consecutive epochs. Additionally, the learning rate was reduced by a factor of 0.1 after every 10 epochs without improvement. The three models stopped their training after 16, 24, and 14 epochs, respectively.

To handle the large imbalance in data, both between AMFs and NMFs, as well as the subclasses, a class-weighted categorical focal loss was employed. The class weights were computed based on the class distributions of the corresponding training subsets of the cross-validation splits.

During training online augmentations like *flip*, *rotate*, and *gaussian noise* were used. During testing no augmentations were applied.

For evaluation, the test images were provided to the three model ensemble for classification and their results were averaged via the mean of their output vectors. Standardized evaluation metrics and visualizations were generated based on the consensus recommendations for medical imaging classification performance measurements [14].

The code as well as evaluation data is available at: https://github.com/hnu-digihealth/GMDS_Mitotic_Figure_Classification.

3. Results

The AUCMEDI trained model achieved varying performance across the different normal and abnormal MFs. Table 1 shows the classification evaluation results for all subtypes on the training data. Overall, a high Specificity (>=93.76%) can be seen. However, the Sensitivity is heterogenous, ranging from 0.0% (AMF Bipolar) up to 82.02% (NMF Metaphase).

Figure 2 shows the confusion matrix for the test data. NMFs show an overall clear separation to AMF. With only two outliners. As the smaller outliners, Ana- & telophase NMF have a 10.5% rate of being classified as bipolar AMF and a 14.47% chance to be classified as NMF Metaphase. Ring shaped NMF being the bigger outliner, being classified to the “other” AMF group 43.75% of the time and 18.75% as Prometaphase NMF. However, no other confusions besides these two exist for Ring shaped NMFs.

With AMFs a more heterogenous image is provided, with all classes showing at least one confusion over 20% and only the "other" AMF group classifying correctly with over 40%.

Table 1. Overview of Area Under the Curve (AUC), Specificity, and Sensitivity for all Classes in the Test Dataset. Further, the Number of Images for each Class in the Test Dataset is Reported.

Class	AUC	Specificity	Sensitivity	Number	PPV	NPV
AMF – Bipolar	54.83 %	97.50 %	0.00 %	21	0.00 %	98.39 %
AMF – Multipolar	94.36 %	99.09 %	13.80 %	29	25 %	98.11 %
AMF - Segregation	82.40 %	95.35 %	15.68 %	51	11.76 %	96.62 %
AMF – Other	89.22 %	94.24 %	43.01 %	93	35.71 %	95.69 %
NMF – Ana- & telophase	89.85 %	93.76 %	60.53 %	79	36.80 %	97.53 %
NMF – Metaphase	93.82 %	90.59 %	82.02 %	534	85.21 %	88.40 %
NMF – Prometaphase	95.26 %	94.76 %	78.16 %	522	90.46	87.20 %
NMF – Ring shape	87.47 %	98.71 %	0.375 %	16	26.09 %	99.24 %

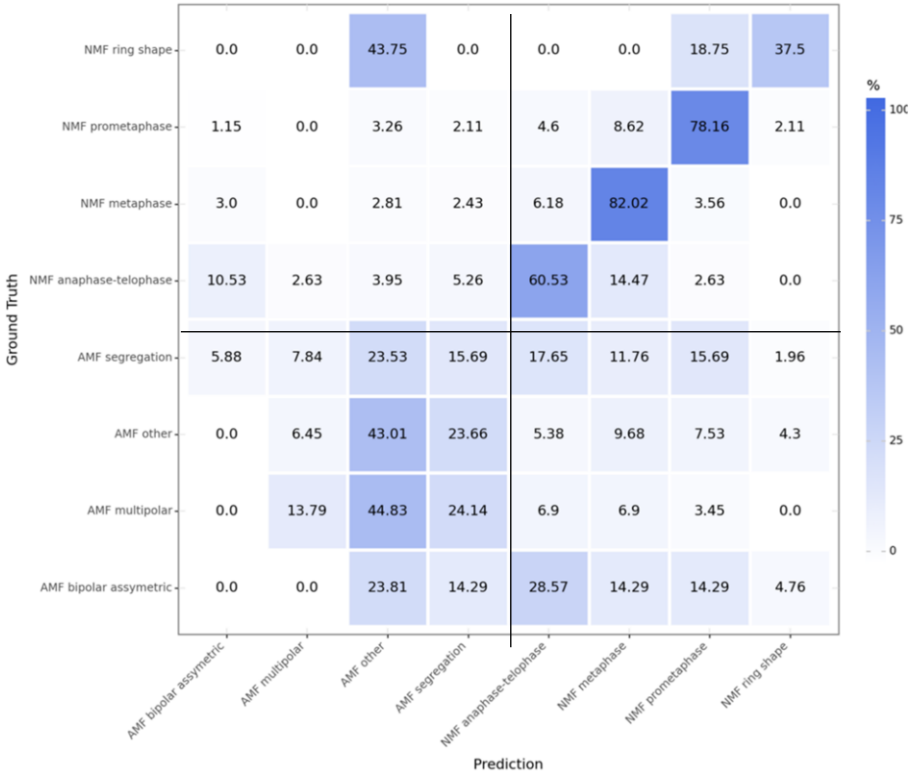

Figure 2. Confusion Matrix for the eight Subclass Classification Problem.

4. Discussion

The overall mean AUC of the eight-subclass classification model achieved 85.90% compared to 69.8% ± 2.6% of the binary classification baseline model by Bertram et al . [7]. Furthermore, we were able to achieve a balanced accuracy of 79% for binary classification in comparison to 71.3% ± 1.6. While this is a significant improvement, the confusion matrix of Figure 1 and the evaluation results in Table 1 show still room for improvement.

While the model overall achieved a good specificity with values over 93% the sensitivity is still extremely heterogenous with values between 0.0% and 82.02%. Overall, a strong degree of correlation (85.34%**) can be seen between the number of available images in the test dataset and the sensitivity for the corresponding class. This could be due to the large range of the task in contrast to the small provide dataset.
Besides the most common solution "increasing the number of available images in the dataset" a model ensemble approach could be used as a first solution approach. As the general distinction between AMF and NMF worked well for most subclasses a binary classification model could be used to first make the binary distinction between AMF and NMF classes and then two specialized subclass classification models (one for AMFs and NMF each) could be used to make the subclass classification. This would significantly reduce the complexity of the specific subtasks compared to a model doing all in one.

Approaches besides model tweaks and performance tuning could be the inclusion of other (multi-domain) datasets of MFs in model training in a semi-supervised or multiple-instance learning approach [15]. Moreover, foundation models could be a promising solution-approach for this task. Some of those (e.g., Virchow [16]) were already trained on MF datasets, which could lead to a solid understanding of the underlaying morphological structures.

Finally, it could be a valid option to build a rule-based approach, using computer vision for object detection, but trying the actual classification with hard coded rules based on general detection guidelines [17].

5. Conclusion

This work highlights a first classification approach for the AMi-Br MF dataset using a deep learning pipeline based on the AUCMEDI framework. With a mean AUC of 85.90%, we demonstrated that deep learning approaches can be a valuable tool for reliable AMF classification. While the presented subclass classification approach was able to surpass the binary baseline results, there is still room for improvements.

Further work on the dataset should focus on dividing the overall problem in smaller subtasks and employing an ensemble based of multiple specialized models for each of those, namely NMF/AMF-distinction and then AMF/NMF subclass classification. Moreover, hyperparameter optimization should be employed to further increase the models' performances.

Declarations

Conflict of Interest: The authors declare no conflict of interest.

Acknowledgement: We would like to thank Marc Aubreville and his whole team for their continues dedication towards open medical data.

Data Availability: The dataset is made publicly available by Bertram, C.A. et al. at https://github.com/DeepMicroscopy/AMI-Br. The analysis code is available at: https://github.com/hnu-digihealth/GMDS_Mitotic_Figure_Classification.

Author Contributions: Conceptualization: DH, DM; Methodology: DM; Software: DH, DM; Validation: DM; Formal Analysis: DM; Investigation: DM; Resources: FL-S, JS, RP, FK; Data Curation: DM; Writing – Original Draft: DH, DM; Writing – Review \& Editing: DH, FL-S, JS, RP, FK, DM; Visualization: DH, DM; Supervision: JS, FK; Project Administration: DM; Funding Acquisition: FL-S, JS, RP, FK;

References

[1] C. van Dooijeweert, P.J. van Diest, and I.O. Ellis, Grading of invasive breast carcinoma: the way forward, *Virchows Arch*. **480** (2022) 33–43. doi:10.1007/s00428-021-03141-2.

[2] C.A. Bertram, T.A. Donovan, and A. Bartel, Mitotic activity: A systematic literature review of the assessment methodology and prognostic value in canine tumors, *Vet Pathol*. **61** (2024) 752–764. doi:10.1177/03009858241239565.

[3] A. Ibrahim, M.S. Toss, S. Makhlouf, I.M. Miligy, F. Minhas, and E.A. Rakha, Improving mitotic cell counting accuracy and efficiency using phosphohistone-H3 (PHH3) antibody counterstained with haematoxylin and eosin as part of breast cancer grading, *Histopathology*. **82** (2023) 393–406. doi:10.1111/his.14837.

[4] Z. Shen, M. Simard, D. Brand, V. Andrei, A. Al-Khader, F. Oumlil, et al., A deep learning framework deploying segment anything to detect pan-cancer mitotic figures from haematoxylin and eosin-stained slides, *Commun Biol*. **7** (2024) 1–11. doi:10.1038/s42003-024-07398-6.

[5] A. Sohail, A. Khan, N. Wahab, A. Zameer, and S. Khan, A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images, *Sci Rep*. **11** (2021) 6215. doi:10.1038/s41598-021-85652-1.

[6] A. Lashen, M.S. Toss, M. Alsaleem, A.R. Green, N.P. Mongan, and E. Rakha, The characteristics and clinical significance of atypical mitosis in breast cancer, *Mod Pathol*. **35** (2022) 1341–1348. doi:10.1038/s41379-022-01080-0.

[7] C.A. Bertram, V. Weiss, T.A. Donovan, S. Banerjee, T. Conrad, J. Ammeling, et al., Histologic Dataset of Normal and Atypical Mitotic Figures on Human Breast Cancer (AMi-Br), in: C. Palm, K. Breininger, T. Deserno, H. Handels, A. Maier, K.H. Maier-Hein, et al. (Eds.), Bildverarbeitung für die Medizin 2025, Springer Fachmedien Wiesbaden, Wiesbaden, 2025: pp. 113–118. doi:10.1007/978-3-658-47422-5_25.

[8] M. Veta, Y.J. Heng, N. Stathonikos, B.E. Bejnordi, F. Beca, T. Wollmann, et al., Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge, *Med Image Anal*. **54** (2019) 111–121. doi:10.1016/j.media.2019.02.012.

[9] M. Aubreville, N. Stathonikos, C.A. Bertram, R. Klopfleisch, N. ter Hoeve, F. Ciompi, et al., Mitosis domain generalization in histopathology images — The MIDOG challenge, *Medical Image Analysis*. **84** (2023) 102699. doi:10.1016/j.media.2022.102699.

[10] M. Aubreville, N. Stathonikos, T.A. Donovan, R. Klopfleisch, J. Ammeling, J. Ganz, et al., Domain generalization across tumor types, laboratories, and species — Insights from the 2022 edition of the Mitosis Domain Generalization Challenge, *Medical Image Analysis*. **94** (2024) 103155. doi:10.1016/j.media.2024.103155.

[11] D. Müller, D. Hartmann, I. Soto-Rey, and F. Kramer, Abstract: AUCMEDI: Von der Insellösung zur einheitlichen und automatischen Klassifizierung von Medizinischen Bildern, in: T.M. Deserno, H. Handels, A. Maier, K. Maier-Hein, C. Palm, and T. Tolxdorff (Eds.), Bildverarbeitung für die Medizin 2023, Springer Fachmedien Wiesbaden, Wiesbaden, 2023: pp. 253–253. doi:10.1007/978-3-658-41657-7_55.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., ImageNet Large Scale Visual Recognition Challenge, *Int J Comput Vis*. **115** (2015) 211–252. doi:10.1007/s11263-015-0816-y.

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, A ConvNet for the 2020s, in: 2022: pp. 11976–11986. https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html (accessed September 8, 2023). doi: 10.48550/ARXIV.2201.03545.

[14] L. Maier-Hein, A. Reinke, P. Godau, M.D. Tizabi, F. Buettner, E. Christodoulou, et al., Metrics reloaded: recommendations for image analysis validation, *Nat Methods*. **21** (2024) 195–212. doi:10.1038/s41592-023-02151-z.

[15] M. Aubreville, F. Wilm, N. Stathonikos, K. Breininger, T.A. Donovan, S. Jabari, et al., A comprehensive multi-domain dataset for mitotic figure detection, *Sci Data*. **10** (2023) 484. doi:10.1038/s41597-023-02327-4.

[16] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, et al., A foundation model for clinical-grade computational pathology and rare cancers detection, *Nat Med*. **30** (2024) 2924–2935. doi:10.1038/s41591-024-03141-0.

[17] T.A. Donovan, F.M. Moore, C.A. Bertram, R. Luong, P. Bolfa, R. Klopfleisch, et al., Mitotic Figures—Normal, Atypical, and Imposters: A Guide to Identification, *Vet Pathol*. **58** (2021) 243–257. doi:10.1177/0300985820980049.