Abstract: P1077

Title: ENHANCING ACCURACY AND SAFETY OF AI-BASED BONE MARROW ASSESSMENT IN MPNS WITH AN AUTOMATED QUALITY CONTROL PIPELINE

Abstract Type: Poster Presentation

Topic: Myeloproliferative neoplasms - Clinical

Background:

The development of AI-based approaches has improved the accuracy and consistency of bone marrow (BM) assessment in myeloproliferative neoplasms (MPN). Effective quality control (QC), mitigating the performance degradation due to the presence of bones and artefacts present in WSIs, thus ensures the safety and reliability of such machine learning (ML) solutions. Here, we introduce an automated QC pipeline to enhance the algorithm performance, and provide a quantitative assessment of the prediction reliability.

Aims:

The aim is to evaluate the impact of QC on ML based assessment of BM. By integrating tissue, bone and artefact masks [Fig b] from the QC results, we aim to improve model performance, and quantify the relationship between analysable tissue area and prediction reliability.

Methods:

We performed a subtype classification task to compare ML models trained with and without QC inputs. We then evaluated prediction reliability by simulating WSIs of various analysable tissue areas as inputs to ML models, mimicking the real-life scenario of using WSIs for diagnosis.

Results:

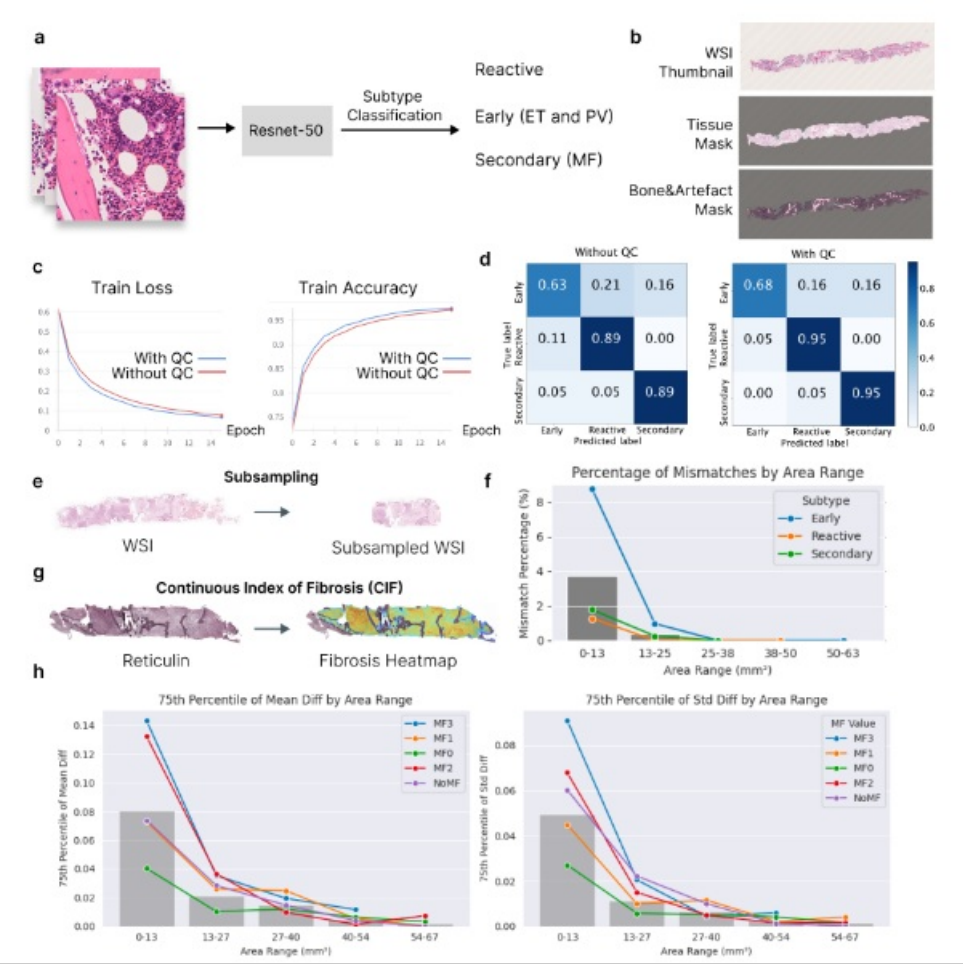
We used 135 hematoxylin and eosin (H&E) stained BM WSIs from INForMeD Study (The University of Oxford) consisting of 45 WSIs in each of the three groups: reactive, early [essential thrombocythemia (ET) and polycythemia vera (PV)] and secondary [myelofibrosis (MF), post-ET MF and post-PV MF]. A ResNet-50 classification model was trained on 512x512 tiles extracted at 20x magnification [Fig a]. We evaluated the impact of QC by conducting two experiments. The first one applied QC masks [Fig (b)] to include tissues while excluding bones and artefacts. The second extracted tissue tiles without discarding the bones and artefacts. The group with QC showed quicker convergence and higher accuracy during training [Fig c]. We then used a separate WSI dataset (n=57) for testing, 19 from each subtype (reactive, early and secondary). During inference, the tile outputs are aggregated to produce a WSI-level prediction by majority vote. The QC group F1 scores and accuracies of 0.855 and 0.868, respectively, against the 0.802 and 0.808 of the group without QC. The confusion matrices reflected better predictions across all three groups [Fig d].

Next, we determined the effect of tissue area on the model's classification performance. On the 57 testing WSIs, 100 tissue areas of variable sizes per WSI [Fig e] are randomly sampled. We evaluated the alignment of the subtype prediction between the sampled and original WSIs. We plotted the percentage of mismatches to the original WSIs [Fig f] against the analysable area, showing tissue areas under 13 mm2 were prone to prediction variability.

Similarly, we assessed Continuous Indexing of Fibrosis (CIF), which is a quantification of reticulin fibrosis (162 reticulin WSIs) [Fig g]. Using the same sampling method, we compared the mean and standard deviation of CIF scores between the subsamples and the full WSIs [Fig h]. Results showed WSIs with less than 13 mm2 of analysable area had unreliable CIF readings. Samples with higher fibrosis grades were more likely to be mismatched upon subsampling if analysable area is limited. Our assessment aligns with WHO guidelines, which recommend a minimum sample length of 15mm (approximately 30 mm2), indicating an analyzable area should exceed 13 mm2.

Summary/Conclusion:

In conclusion, our QC pipeline not only elevates ML model performance, but also establishes a relationship between analysable area and the prediction reliability. This approach should be used in future ML based WSI solutions, for enhancing model performance and understanding model limitations.



Keywords: Myelofibrosis, Myeloproliferative disorder, Artificial intelligence, Diagnosis