An approach to resource saving histology dataset expansion

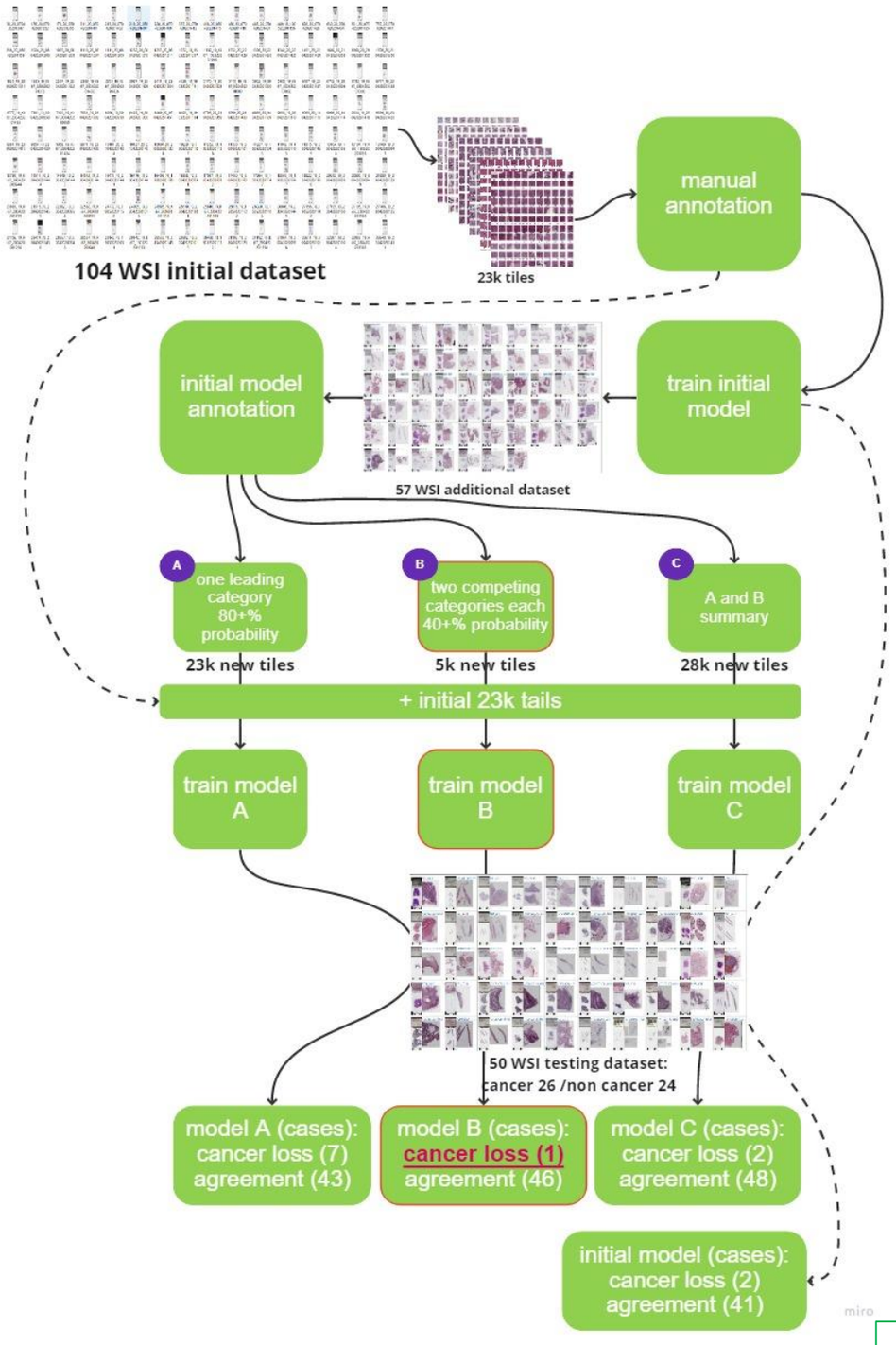
Artyom Borbat[1,2], Inna Yatsenko[2]
1. Leycor LLC (Moscow, Russia) 2. Burnasyan Federal Medical Biophysical Center Of Federal Medical Biological Agency (Moscow, Russia)
email: aborbat@yandex.ru

Introduction

Annotated datasets are still a bottle neck for artificial intelligence in pathology. We report an approach to expand existing histology datasets with less efforts from pathologists.

Materials and methods

Previously published dataset of breast lesions (104 cases, 23 k tiles), 57 new cases to expand dataset and 50 new test cases. Classification groups: non-specific invasive cancer differentiated, non-specific invasive cancer low-differentiated, lobular invasive cancer, non-invasive ductal carcinoma, fibroadenoma, fibrocystic changes, papilloma and background. Initial convolutional neural network model was trained with published dataset and applied to new cases on a tile-by-tile bases. Two kinds of resulting tiles were captured for further steps. Confident tiles with classification probability 80% and more (overall 46 k tiles). And doubt tiles, which included two competing categories, each not less than 40% probability (overall 28 k tiles). The third group was a summary of two groups (overall 50 k tiles). These groups were added to published dataset. Based on the expanded datasets three new models were trained and tested with test cases. The analysis results were evaluated by a pathologist.



Results

We identified, that models trained on dataset with doubt tiles outperformed initial model and model with confident tiles in malignant vs benign tissue, though the quantity of doubt images added less than 20% to initial published dataset.

Conclusion

To form a new training dataset based on existing one, one should include controversial or imperfect tiles. It improves model learning even with fewer image quantities.

- ✓ **initial dataset can be expanded by annotating with initial model and specified probability**
- ✓ **train datasets (B and C) with controversial tiles provided better models' performance**
- ✓ **even with smaller dataset size (B)**

The research is granted by
Foundation for Assistance to Small Innovative Enterprises (FASIE)
119034, 1/5, 3rd Obydenskiy per, Moscow, Russia