HISTOPATHOLOGY BREAST CANCER IMAGE CLASSIFICATION USING SUPERVISED LEARNING

Natalia Gil Canto - 232881

Sadeeq Olalekan Bello - 198346

Suelen Aparecida Ribeiro de Souza - 252483

Temitope Kola Adebowale - 256247

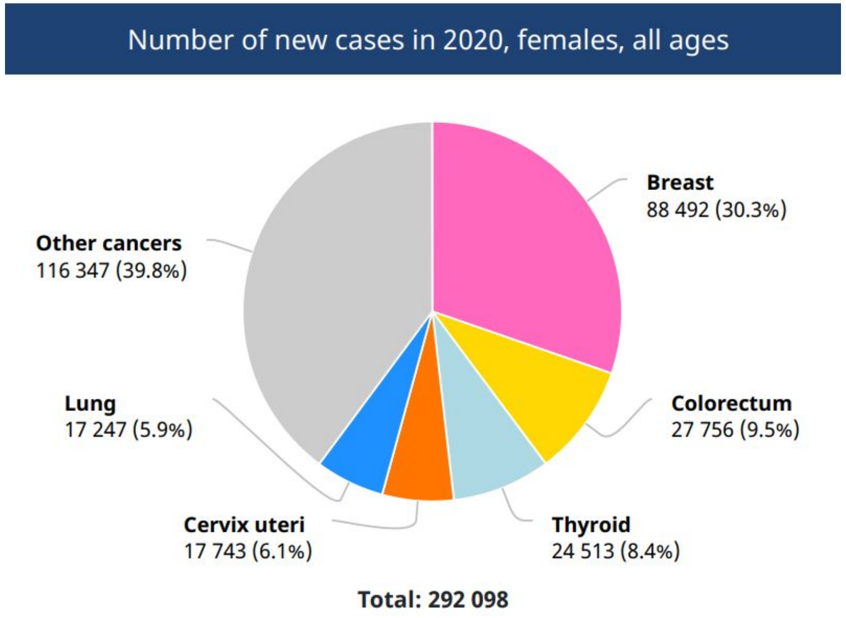
Health Data Science and Visualization

Agenda

- Introduction
- Objectives and research question
- Methodology
- Results and Discussion
- Conclusions
- Future work
- References

Introduction

- Breast cancer is one of the leading causes of death among women in Brazil. Most cases, 80% to 90% of them, represent invasive ductal carcinoma (IDC). Early-stage diagnosis: approximately 95% chance of cure for patients.
- If there are indications of a tumor or significant abnormalities, a biopsy is performed, in which a small amount of tissue is collected. The tissue undergoes a process that includes fixation, embedding, sectioning, mounting on slides, and staining with hematoxylin and eosin.
- To analyze the slides, it is necessary to manually identify the affected or altered regions, an activity manually made by pathologists.



- With advances in research and the use of technologies, it is expected that breast cancer detection will become increasingly early, allowing the use of targeted treatments and thus saving more lives (Aksac et al., 2019; NIH, 2022).
- Artificial intelligence is an option because it could assist pathologists on reducing the need for laborious manual analysis of the tissue slides. With that, the pathologist could focus especially on the validation of the technique.

Objectives and research question

- **Research Question:** Can supervised learning be used to identify specific features or patterns in histopathology images that are indicative of breast cancer?

- **General objective:** To evaluate the effectiveness of a supervised learning-based image classification model in detecting breast cancer on histopathology images.

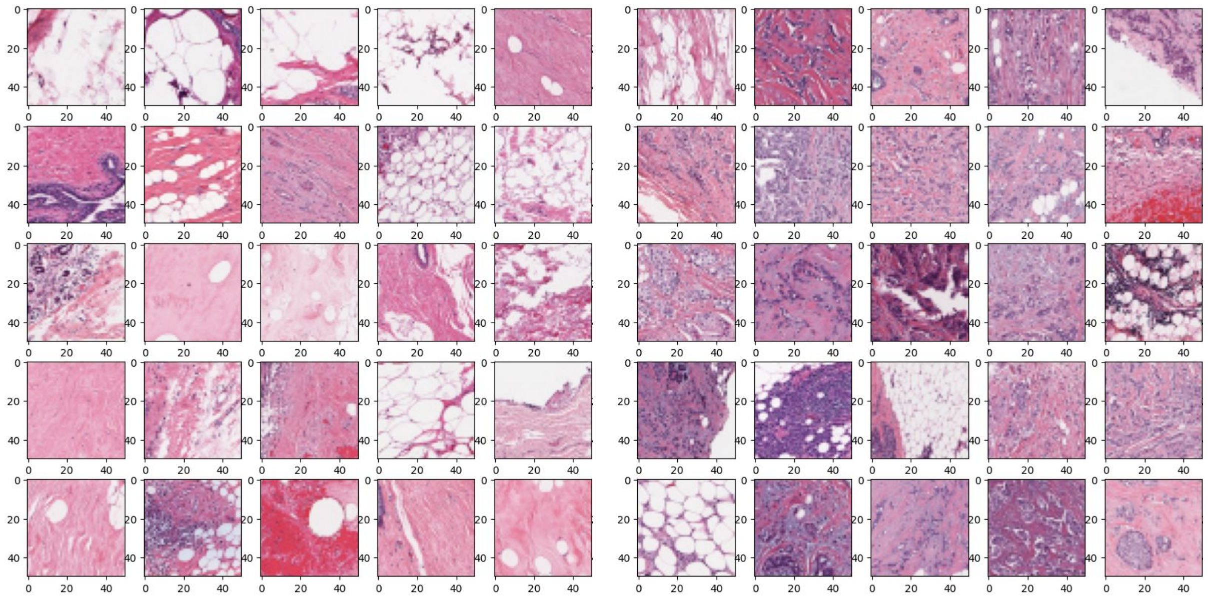
- **Specific objectives**

1. To evaluate the performance of an image classification model on histopathology images using different metrics: accuracy, recall, specificity, f1-score and precision;
2. To analyze the model's incorrect and correct classifications to get insights on its behavior;
3. Propose improvements on the model's learning process and future work.

Methodology

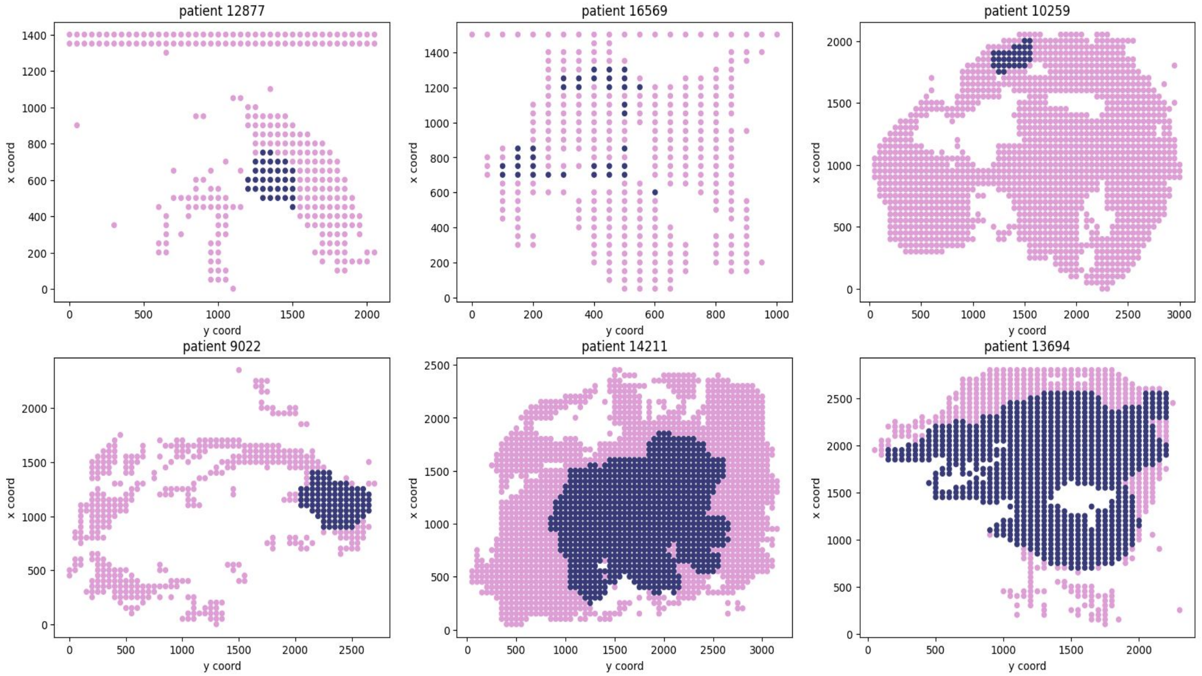
- The dataset

The dataset was achieved through Whole Slide Imaging (WSI), which is a technology that digitizes entire histological or cytological tissue slides, allowing high-resolution imaging of the entire glass slide. It replaces conventional microscopy by capturing detailed images of the tissue or cells.



Breast Histopathology Images

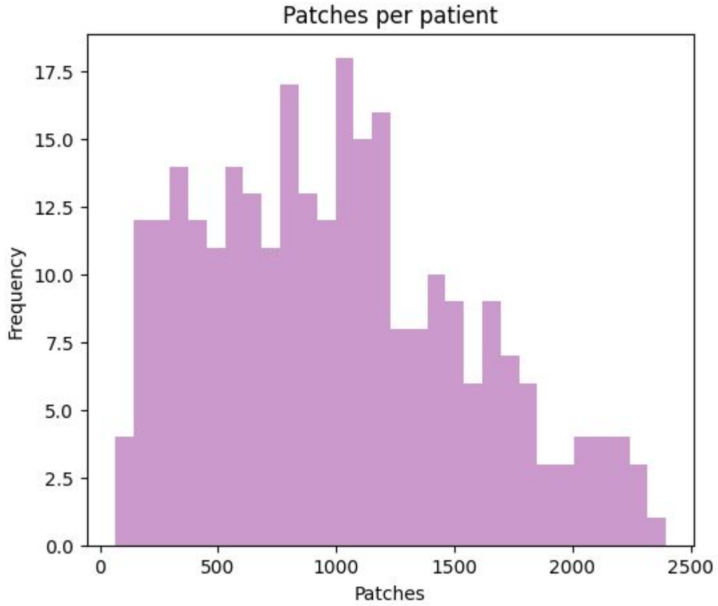
- 279 patients
- 277,524 patches of size 50 x 50 pixels were extracted
- 198,738 IDC negative
- 78,786 IDC positive



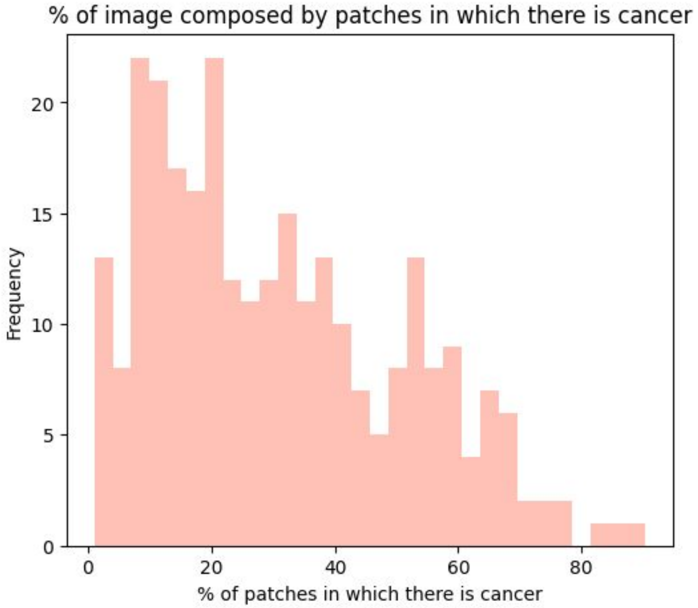
Based on the patient's patches, it is possible to reconstruct the tissue.

Methodology

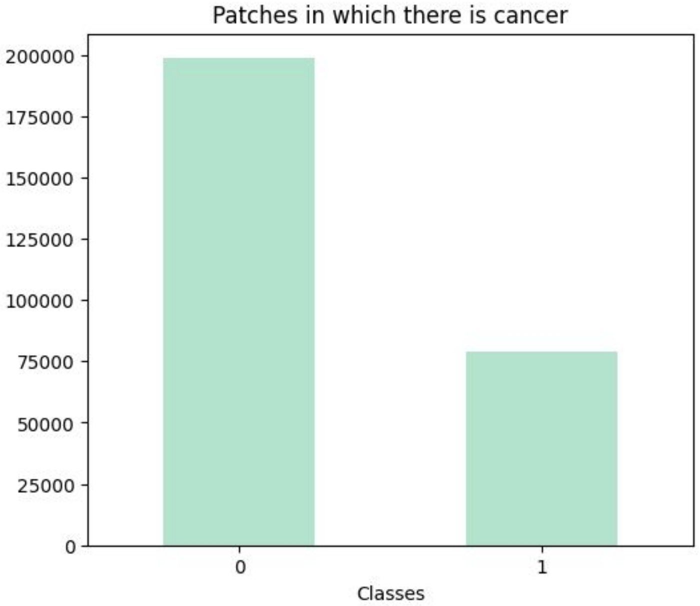
- Exploratory Analysis



If the patches are the same size (50x50) per patient and the total number is still different, maybe the full images per patient have different sizes or some data is missing.



Around 20% of the tissues are composed by patches in which IDC are identified.



Imbalanced classes. It may produce difficulties on the learning of the positive class.

Methodology

- The classifier
  - EfficientNetV2S
    - Image classification model established in the literature;
    - Promises a balance between accuracy and training time;
    - Usually performs well on mid-sized and high-sized datasets;
    - Available on Keras library.
- Preprocessing
  - Data split into train (70%), validation (15%) and test (15%) sets.
  - Data split by unique patients to not mix patient patches between different sets.
  - Data truncating to achieve balance between classes on each set.
  - Data reduction by 50% due to computational constrains.

Before preprocessing:

Train size: class0-142774 | class1-55437
Validation size: class0-23903 | class1-9296
Test size: class0-32061 | class1-14053

After preprocessing:

Train size: class0-27718 | class1-27718
Validation size: class0-4648 | class1-4648
Test size: class0-7026 | class1-7026

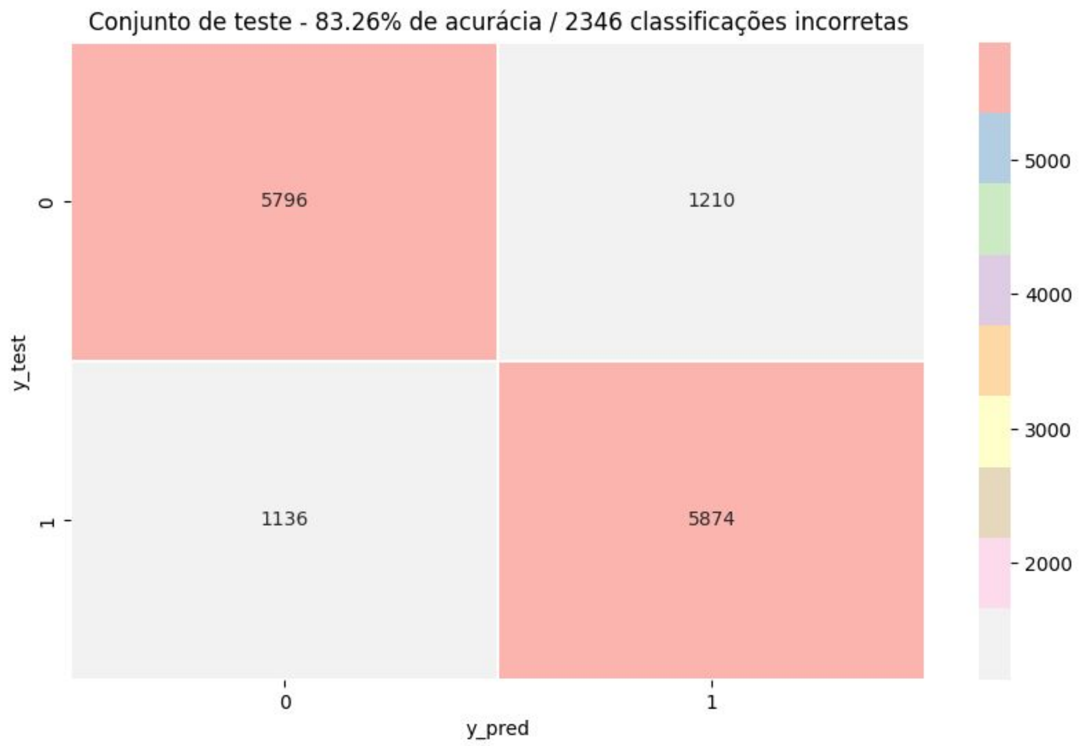
Results and discussion

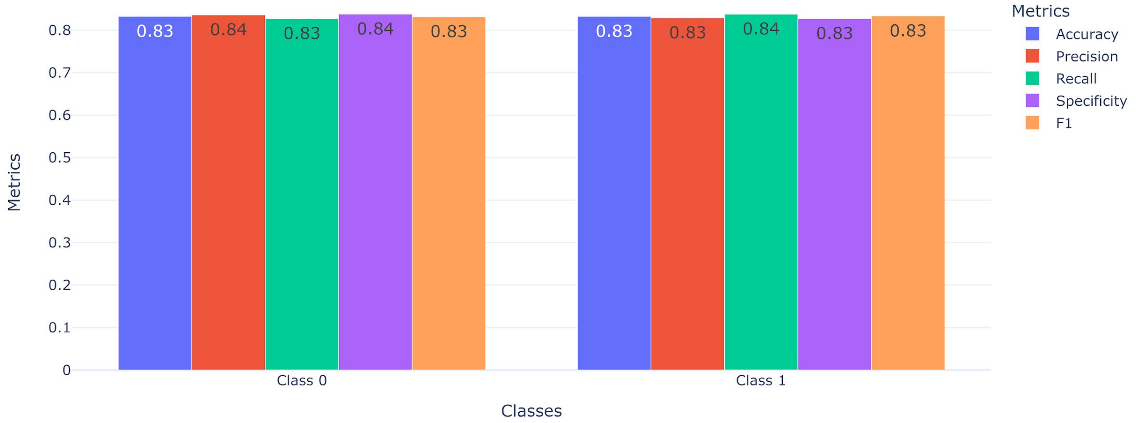
In the medical context, evaluating the false negatives produced by the classifier is crucial as they represent instances where the model failed to identify breast cancer in the tissue despite its presence.

Considering the overall sample size of the test set, the model demonstrated a certain level of proficiency in learning the primary patterns indicative of breast cancer.

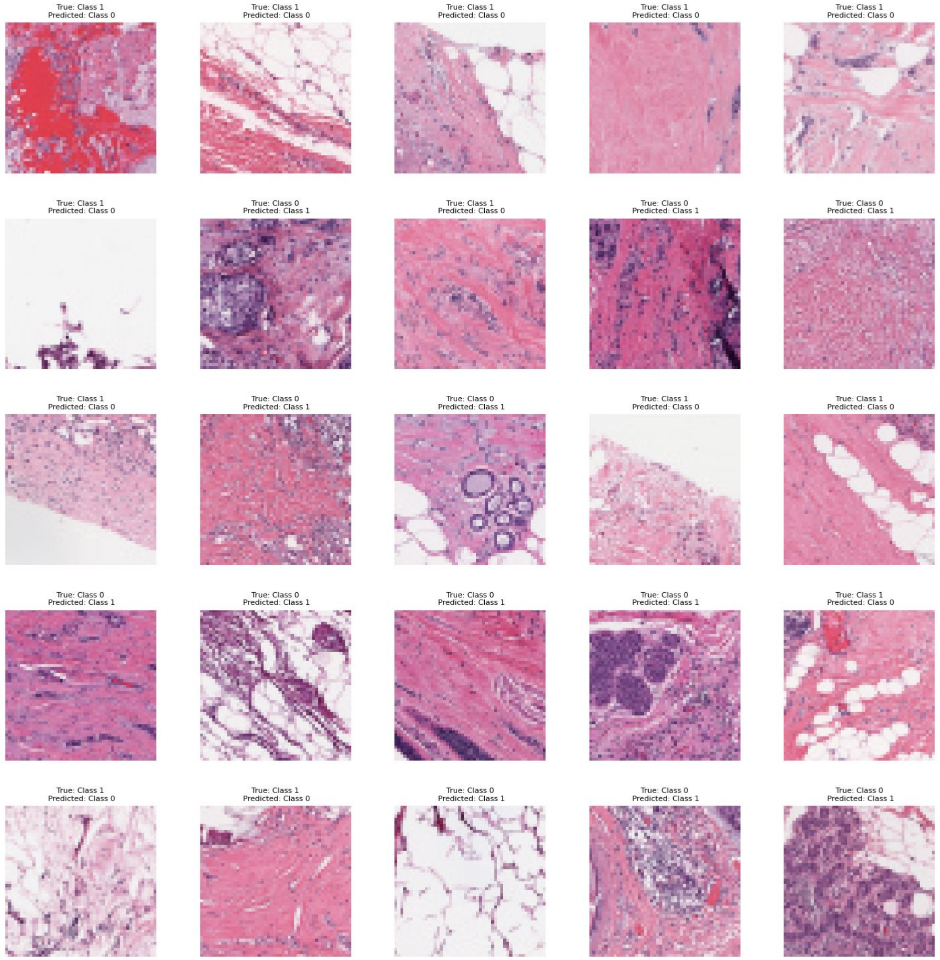
The precision, recall, specificity, accuracy and f1-score present similar behavior, indicating the model presents consistent performance metrics around 83% for both classes, indicating an interesting level of effectiveness in classifying breast cancer tissues on the given dataset.

We do believe this consistent performance in multiple metrics is, also, a result of the data balance performed in the preprocessing stage.

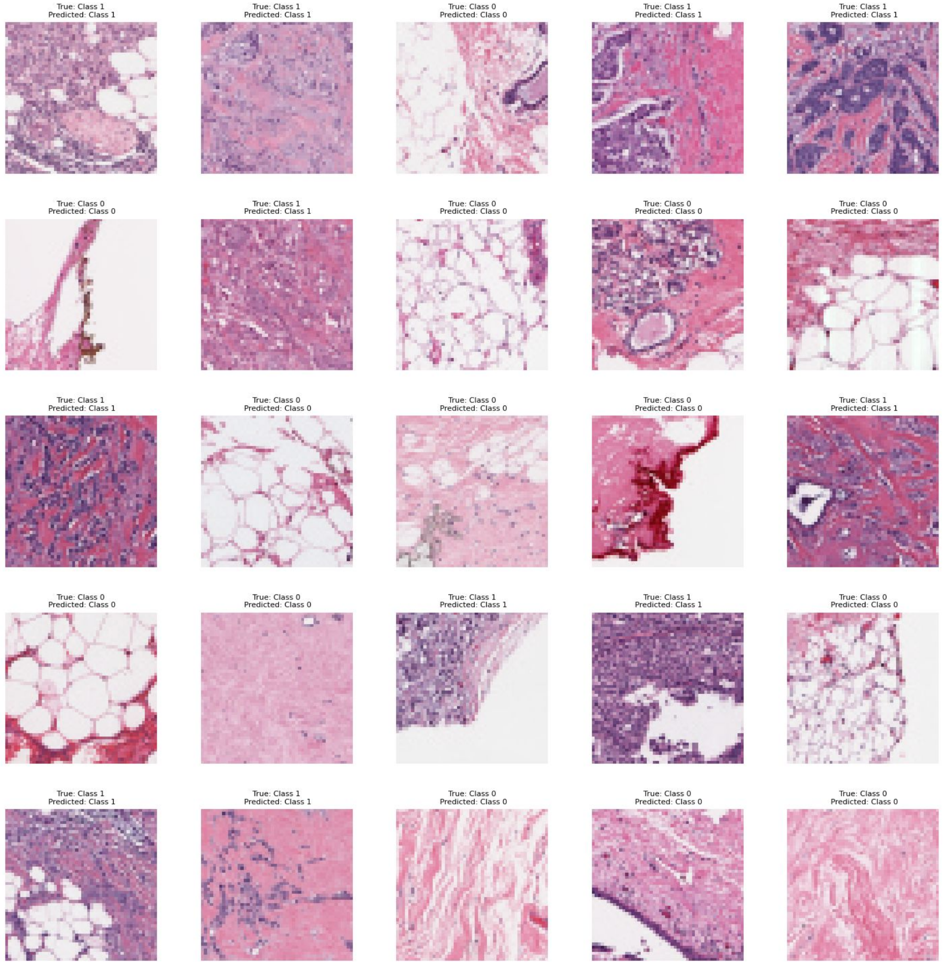




Results and discussion



Examples of the incorrect classifications



Examples of the correct classifications

Conclusions

- From the results and discussions we have made, we can positively answer the research question: can supervised learning be used to identify specific features or patterns in histopathology images that are indicative of breast cancer?

- We have observed, even though we are not using robust techniques nor exhaustive training, the classification model was capable of successfully concluding the desired task required.

- The presence of a qualified professional in the field is still highly important. This is because we are uncertain about the methods that the model is using when making decisions.

- Also, the use of these techniques would not replace the work of pathologists, but rather shorten the time required for the manual analysis steps they need to perform.

- Furthermore, the final evaluation carried out by professionals is crucial to distinguish real cancer features from variations related to employed histological techniques and other subjectivities in image interpretation.

Future work

- What could we have done if we have more time?

  - Comparing different classification models;
  - Increasing the number of training epochs;
  - Using transfer learning to improve generalization;
  - Explore segmentation techniques to highlight specific areas;
  - Apply the trained/validated model on other types of tissues.

References

Abrahamsohn, Paulo; Freitas, Vanessa. Histology: 1-23 Basic Concepts. Cellular and Development - Institute of Biomedical Sciences, University of São Paulo, 2004. Available at: https://mol.icb.usp.br/index.php/1-23-conceitos-basicos/.

Aksac A, Demetrick DJ, Ozyer T, et al. BreCaHAD: A dataset for breast cancer histopathological annotation and diagnosis. BMC Res Notes.2019;12:82.

Aquino, R. G. F. D., Vasques, P. H. D., Cavalcante, D. I. M., Oliveira, A. L. D. S., Oliveira, B. M. K. D., & Pinheiro, L. G. P. (2017). Invasive ductal carcinoma: the relationship between pathological characteristics and the presence of axillary metastasis in 220 cases. Rev. Col. Bras. Cir., 44(2), 163–170. doi:10.1590/0100-69912017002010.

Bolhasani H, Amjadi E, Tabatabaeian M, Jassbi SJ. A histopathological image dataset for grading breast invasive ductal carcinomas. Informatics in Medicine Unlocked. 2020;19:100341.

Brancati N, Anniciello AM, Pati P, Riccio D, Scognamiglio G, Jaume G, De Pietro G, Di Bonito M, Foncubierta A, Botti G, Gabrani M, Feroce F, Frucci M. BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. Database (Oxford). 2022 Oct 17;2022:baac093. doi: 10.1093/database/baac093. PMID: 36251776; PMCID: PMC9575967.

Couture HD, Williams LA, Geradts J, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype and intrinsic subtype. NPJ Breast Cancer.2018;4:30.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., … Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Medical Imaging 2014: Digital Pathology. doi:10.1117/12.2043872

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).

References

Jaroensri R, Wulczyn E, Hegde N, et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. NPJ Breast Cancer. 2022;8:113.

Kaggle. Breast Histopathology Images. 2018. Available at: https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images. Accessed April 28, 2023.

Mercan C, Balkenhol M, Salgado R, et al. Deep learning for fully-automated nuclear pleomorphism scoring in breast cancer. NPJ Breast Cancer. 2022;8:120.

Ministry of Health. Concept and Magnitude. Brazil: INCA; 2022. Available at: https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/conceito-e-magnitude. Accessed May 6, 2023.

National Cancer Institute. United States Government: NIH; 2022. Available at: https://www.cancer.gov/types/breast/research. Accessed May 6, 2023.

UNICAMP. Department of Pathology Anatomy (Anatpat). Accessed on May 16, 2023. Available at: anatpat.unicamp.br.

Thank you!