AI Health 2024, Athens, 10-14 March 2024

# CAD Tool for Breast Cancer Prediction using Multiple Deep-learning Models


***Maura Mengoni, Polytechnic University of Marche***
*(m.mengoni@univpm.it)*

***Aubudukaiyoumu Talipu, Giampiero Cimini, Marco Luciani, Luca Giraldi, EMOJ Srl***
*(a.talipu; g.cimini; m.luciani; l.giraldi@emojlab.com)*

***Mauro Savino, Diatech Pharmacogenetics Srl***
*(mauro.savino@diatechpharmacogenetics.com)*


UNIVERSITÀ POLITECNICA DELLE MARCHE

DIISM
Dipartimento di Ingegneria Industriale e Scienze Matematiche

EMOJ
emojlab.com

diatech pharmacogenetics

# PRESENTATION OUTLINE

Introduction

Background

Methodology

Experiments and Results

Conclusion

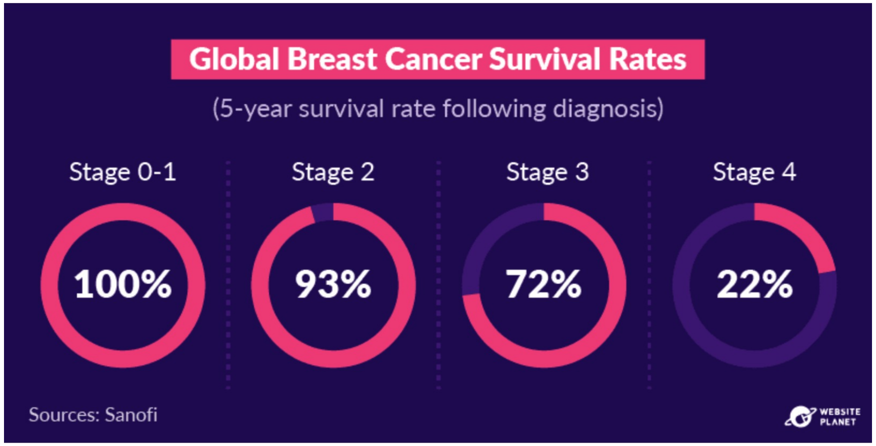
Future work

# INTRODUCTION



Breast cancer is the most common type of cancer worldwide and the leading cause of death among women [1]. Early detection of breast cancer is in vital importance because it is more treatable and gives the patient higher chance of survival.

Breast cancer recurrence is also a crutial field of research, about 80% of patients initially presenting with early-stage disease have a 5 year recurrence, and 30% of patients have a recurrence in 10 years after the complition of initial treatment [2].




[1] "The Global Breast Cancer Initiative." Accessed: 01. 2024. [Online]. Available: https://www.who.int/initiatives/global-breast- cancer-initiative
[2] C. Mazo, C. Aura, A. Rahman, W. M. Gallagher, and C. Mooney, "Application of Artificial Intelligence Techniques to Predict Risk of Recurrence of Breast Cancer: A Systematic Review," *J Pers Med*, vol. 12, no. 9, pp. 1-11, 2022, doi: 10.3390/jpm12091496.

# BACKGROUND



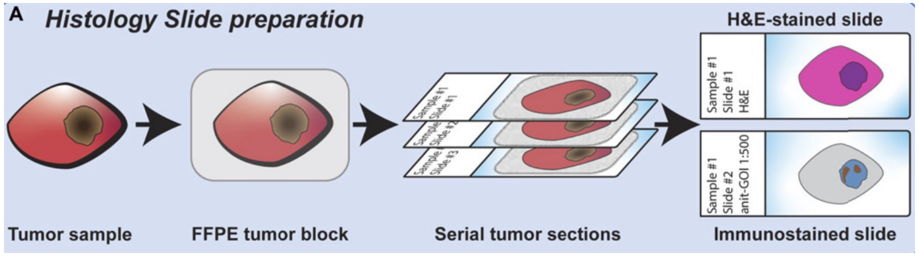
Histopathological tissue analysis by a pathologist is the only definitive method for diagnosing breast cancer (positive/negative) and grading it. (measurement of disease progression)

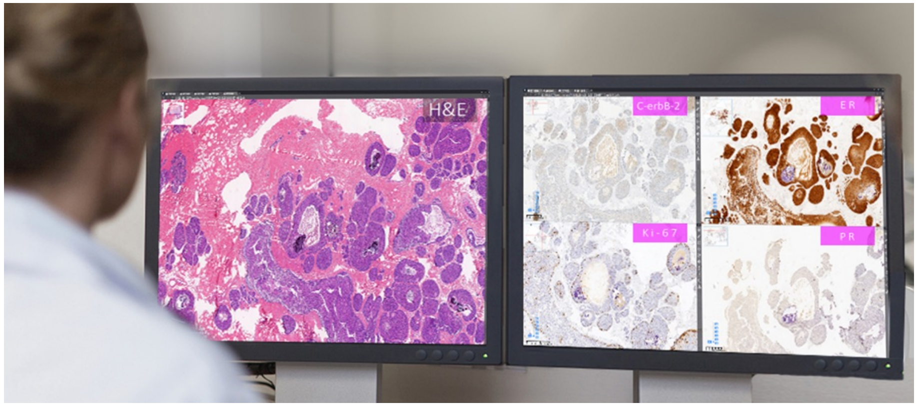
It is a tedius and subjective, inter-observer variation exists event among senior pathologists.

Computer-Aided Diagnostics (CAD) systems can overcome these difficulties, what's more, it can also detect patterns not visible to human eye. [3]

Conventional machine learning based on hand-engineered features [4].

DCNN based machine learning learn useful features directly from training images.




[3] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," *Translational Research*, vol. 194. pp. 20 2018. doi: 10.1016/j.trsl.2017.10.010.

[4] Lee, Kyubum, et al. "Deep learning of histopathology images at the single cell level." Frontiers in artificial intelligence 4 (2021): 754641.

# CAD TOOLS



- Aiforia
- PathAI

Similar features such as automated image analysis, quantification of pathology features and pattern recognition.

- Adjuvant!
- CanAssist-Breast(CAB)
- PREDICT

**Adjuvant!**

- Low accuracy
- No longer available online [4]

**CanAssist-Breast(CAB)**

It utelizes biomarkers and clinical parameters such as tumor size and generates risk score for 5 year recurrence.

**PREDICT**

Effective on 5 year recurrence prediction slight overestimate for 10 year prediction.

[4] N. A. De Glas *et al.*, "Validity of adjuvant! Online program in older patients with breast cancer: A population-based study," *Lancet Oncol*, vol. 15, no. 7, pp. 1, 2014, doi: 10.1016/S1470-2045(14)70200- 1.

# PROPOSED METHODOLOGY



Current existing CAD tools perform only one of the different stages of the diagnostic process; obtaining a general all-in-one diagnostic report requires the involvement of different tools, which is not an efficient workflow.

The primary purpose of the study is to create a CAD tool capable of:

- Classification of cancerous and non-cancerous areas
- Prediction of cancer stage
- A generalised estimation of 10 year breast cancer recurrence

The approach is based on the combination of histological and clinical patient data.

# PROPOSED METHODOLOGY



The development of the novel CAD tool consists of the following steps:

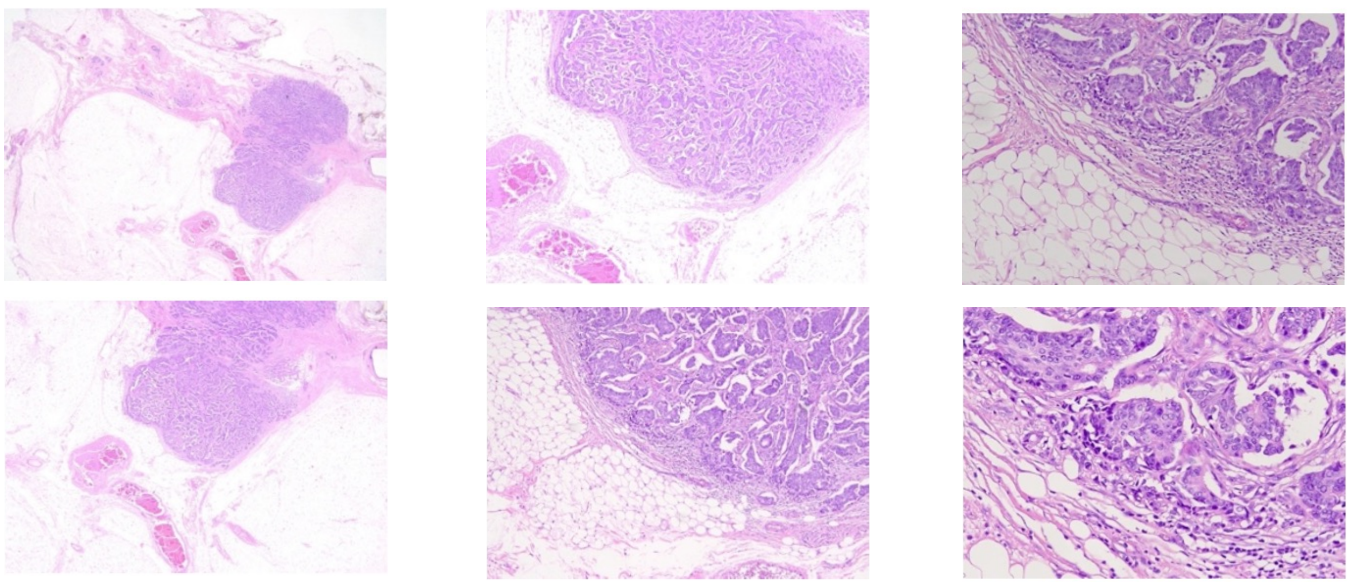
1. Data collection
   - Histological image labelling
2. Dataset creation
   - Stain normalization
   - Patch extraction
3. Model training and validation
   - Supervised deep convolutional neural networks (DCNN)
   - Extream gredient boosting (XGBoost)
   - Linear regression

# DATA COLLECTION

By digitizing 300 samples collected from anonymous patient biopsy slides provided by Verona Borgo Trento Hospital (Italy) with NED DP digital microscope.

Different magnification samples are taken such as 1x, 1.25x, 2x, 4x, 10x, 20x, 40x.

Image resolution is 1640x1175.

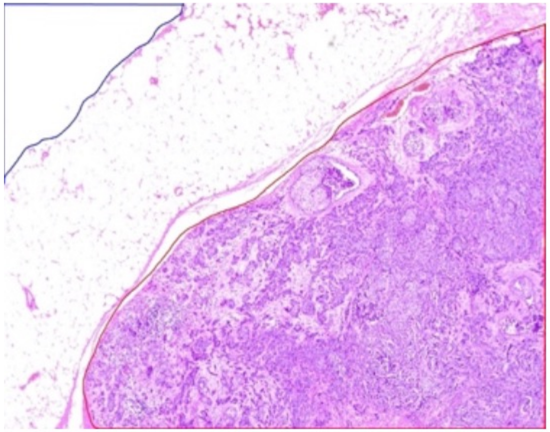


# DATA COLLECTION



Digitized histological images are labelled by the hospital medical practitioners.

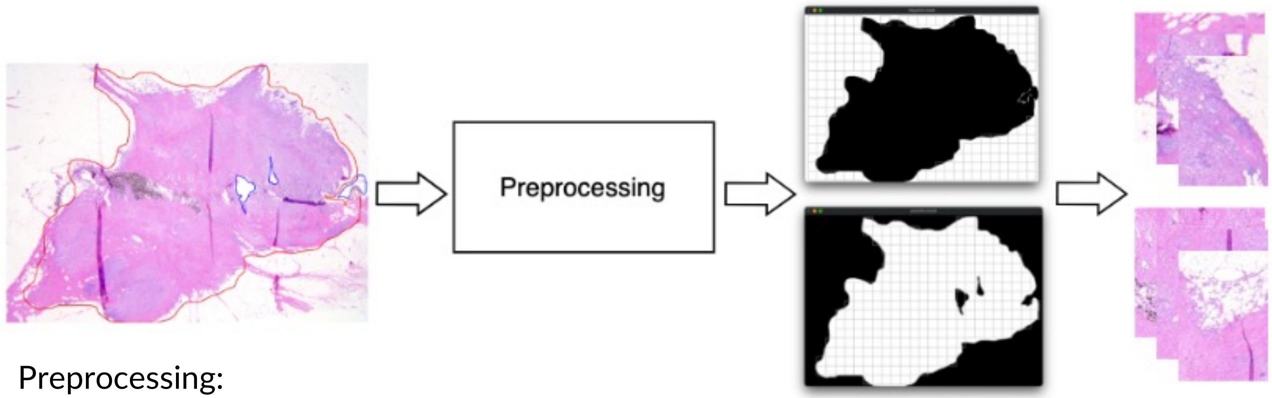
Each sample image also accompanied by corresponding clinical information and features



| All Features | | |
|---|---|---|
| pT | HER2 | RECIDIVA |
| Numero LN metastici | PR | tipo di recidiva |
| pN | N NPI | tempo di recidiva (mesi) |
| Grado | NPI SCORE | Follow up mesi |
| STADIO | NPI GROUP | Luminal |
| DIAMETRO MM | Adiuvante | Età alla diagnosi |
| ISTOTIPO | Overall survival (mesi) | Menopausa |
| Ki67 | DOA | ER |

# DATASET CREATION AND MODEL TRAINING

Separating labels with OpenCV [1]
Stain normalization [2]
Multiple datasets with various patch dimensions 64x64, 150x150 and 200x200.




Preprocessing:
- Stain normalization [5]
- Label separation

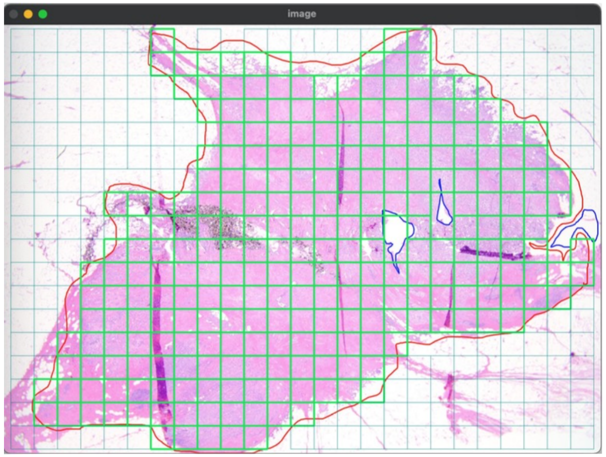
Patch dimensions:
- 64x64
- 150x150
- 200x200

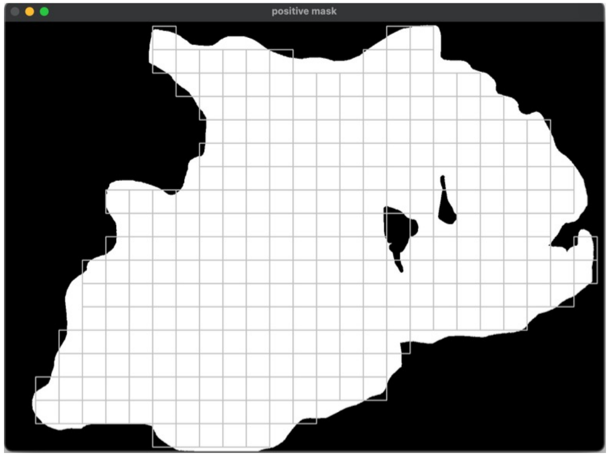
[1] OpenCV computer vision library, Available on " https://opencv.org ".
[2] N. A. De Glas *et al.*, "Validity of adjuvant! Online program in older patients with breast cancer: A population-based study," *Lancet Oncol*, vol. 15, no. 7, pp. 1, 2014, doi: 10.1016/S1470-2045(14)70200- 1.

# SEPARATING LABELS WITH OpenCV



- Using HSV color spectrum values the color-coded labels are separated:
  - Area without tissue
  - Area with cancerous cells
- Areas without tissue are excluded
- 0.7 threshold is applied to selecting positive patches



# DATASET CREATION AND MODEL TRAINING



- Multiple datasets are created based on magnification levels and patch dimensions.
- Multiple DCNN architectures are experimented with the datasets to obtain the best result.
- Fine tuning ImageNet pretrained DCNN models are also experimented.
- Hyper parameters applied during training:
    - Batch size
    - Learning rate

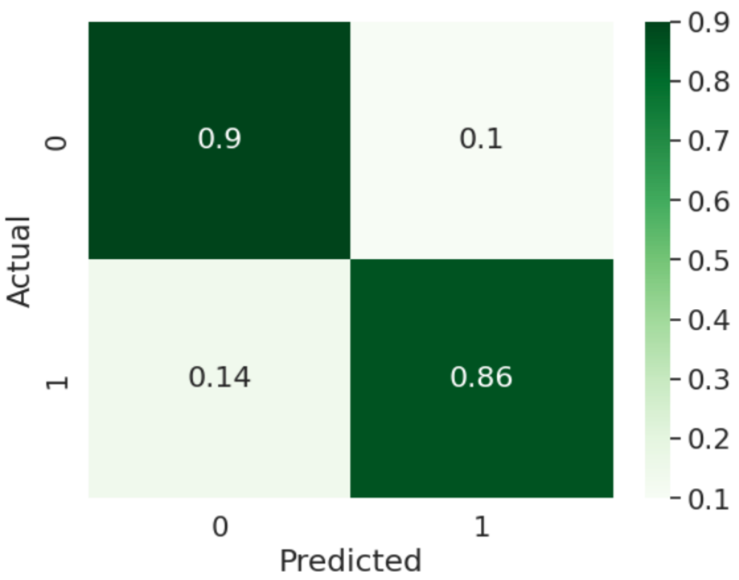
| **Architectures** | VGG16 | VGG19 | Inception |
|---|---|---|---|
| **Patch dimensions** | 64x64 | 150x150 | 200x200 |
| **Batch size** | 32 | 64 | 128 |
| **Learning rate** | 0.01 | 0.001 | 0.003 |

# EXPERIMENT AND RESULTS



- Histological image classification
  - Fine tuning VGG16 model using 40x magnification and 200x200 patch size with 128 batch size and 0.001 learning rate resulted the best accuracy model.
  - Confusion matrix obtained on testing dataset.

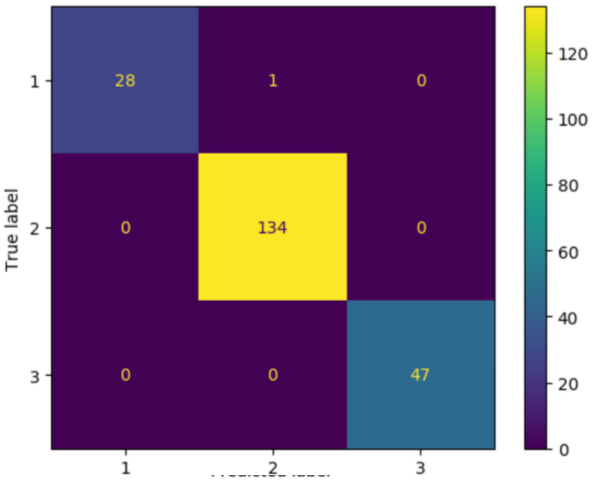
| | |
|---|---|
| **Architecture** | VGG16 |
| **Magnification** | 40 |
| **Patch dimension** | 200x200 |
| **Batch size** | 128 |
| **Learning rate** | 0.001 |
| **Accuracy** | 87.6% |
| **F1** | 0.88 |

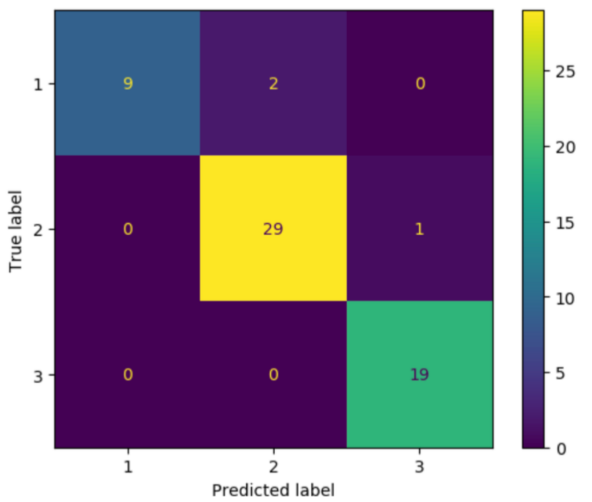


# EXPERIMENT AND RESULTS

- ❑ Cancer grade prediction
  - Following features selected and used in XGBoost algorithm to train the grade prediction model.
  - Confusion matrix obtained on testing dataset

| | | |
|---|---|---|
| Grado | ISTOTIPO_APOCRINO | ADIUVANTE_CHT |
| pT_adjusted | ISTOTIPO_LOBULARE | ADIUVANTE_RT |
| Numero LN metastatici_adjusted | ISTOTIPO_TUBULARE | ADIUVANTE_CT |
| pN_adjusted | ISTOTIPO_NST | ADIUVANTE_OT |
| STADIO_adjusted | ISTOTIPO_CLI | Follow up mesi |
| DIAMETRO MM | ISTOTIPO_CDI | LUMINAL_adjusted |
| ISTOTIPO_PAPINCAPS | Ki67 | età alla diagnosi |
| ISTOTIPO_CRIBRI | PR | menopausa |
| ISTOTIPO_MUCINOSO | N NPI | ER |
| ISTOTIPO_MICROPAPILLARE | NPI SCORE | % cellule neoplastiche |



# EXPERIMENT AND RESULTS



- Recurrence prediction

Because of the ambiguity in the dataset recurrence months, many rows are discarded, and the multiple regression model is trained with very few data, around 40 records.

The final accuracy obtained with the multiple linear regression for 10 - year recurrence prediction is 71%.

# CONCLUSION



The work presented in this paper aims to create an all-in- one breast cancer diagnostic tool for (ER)-positive breast cancer patients.

The histological image classification by fine- tuning ImageNet pretrained VGG16 model obtained 88% accuracy.

The cancer grade prediction with XGBoost algorithm achieved 95% accuracy.

The cancer recurrence prediction with linear regression resulted 71% accuracy.

Histological image analysis and clinical data analysis are combined in the proposed CAD tool to predict breast cancer recurrence. This type of CAD tool is very useful in assisting doctors to reduce their workload and improve the reproducibility of breast cancer diagnostics.

# FUTURE WORK



Further studies are required to improve the accuracies and robustness of the models, this is an initial step in our future study direction.

Exploring integration of molecular and genetic data to enhance the diagnostic accuracy of the deep learning models.

Fusing these heterogeneous data sources might also enable a comprehensive understanding of disease characteristics.