CAD Tool for Breast Cancer Prediction Using Multiple Deep-learning Models

Maura Mengoni
Department of Industrial Engineering and Mathematical Science
Polytechnic University of Marche,
Ancona, Italy
e-mail: m.mengoni@univpm.it

Aubudukaiyoumu Talipu, Giampiero Cimini, Marco Luciani
Technical Department
EMOJ S.r.l.
Ancona, Italy
e-mail: a.talipu@emojlab.com, g.cimini@emojlab.com, m.luciani@emojlab.com

Luca Giraldi
Department of Economy
University of Macerata,
Macerata, Italy
e-mail: luca.giraldi@unimc.it

Mauro Savino
Research and Development
Diatech Pharmacogenetics S.r.l.
Jesi, Italy
e-mail: mauro.savino@diatechpharmacogenetics.com

***Abstract*—Breast cancer is one of the leading causes of cancer death among women worldwide. It represents a global health concern due to the lack of effective therapeutic regimens that could be applied to all breast cancer patients. Breast cancer treatment decisions rely on clinicopathologic parameters. However, this approach is replete with limitations as it fails to define prognosis uniquely and is not always sufficient to settle unequivocally on the best type of treatment for breast cancer patients. The molecular diagnostic efforts have been focused mainly on Estrogen Receptor (ER)-positive (Luminal A) breast cancer being the most represented breast cancer subtype (70% of patients) with a standard treatment (endocrine therapy for five years) and a good prognosis. However, at least 20% of patients will suffer a distant recurrence within ten years. Although many molecular tests have been developed to identify the patients at risk of recurrence, a definite, reliable and effective in vitro diagnostic device that stratifies patients at high risk and low risk of relapse, directing therapeutic decisions, is still a significant clinical need. This study aims to fill this gap by investigating and developing a new approach for better stratification of breast cancer patients in the risk categories of recurrence. It is based on the integration of clinical and digital pathology analysis. The combined analysis, indeed, aims to further categorize the patients with an intermediate risk of recurrence either in the low-risk group with no necessity of chemotherapy or in the high-risk group that needs chemotherapy. The paper presents the approach, the implemented Computer-Aided Diagnosis (CAD) tool and finally, the results of evaluating its predictive accuracy. The tool achieved 88% accuracy in histological image classification, 95% in cancer grade prediction and 71% in 10-year recurrence prediction.**

Keywords: breast cancer; Computer-Aided Diagnosis; Histopathological Imaging; Artificial Intelligence

I. INTRODUCTION

Breast cancer is the most common type of cancer worldwide and the leading cause of death among women [1]. It is worth noticing that early detection and timely diagnosis of breast cancer are of vital importance in saving lives.

Cancer screening helps detect cancer or precancerous abnormalities in individuals with no symptoms. The primary goal of cancer screening is to identify cancer at an early stage when it is more treatable, potentially leading to better outcomes and increased chances of survival. Currently, histopathological tissue analysis by a pathologist represents the only definitive method for confirmation of the presence or absence of disease and disease grading or the measurement of disease progression [2].

Histopathology slides provide a comprehensive view of disease and its effect on tissues because the preparation process preserves the underlying tissue architecture. As such, some disease characteristics may be deduced only from a histological image. However, the histological image analysis process is tedious and subjective, causing inter-observer variations even among senior pathologists [3]. With the advancement of computer vision and image processing based on deep learning algorithms, Computer-Aided Diagnostic (CAD) systems can overcome these difficulties. It can extract the essential information from the histological images and detect patterns not visible to the human eye [4].

Another crucial field of research area related to breast cancer is the prediction of breast cancer recurrence. About 80% of patients initially presenting with early-stage disease have a recurrence in 5 years, and 30% of patients have a recurrence of cancer within 10 years after the completion of initial treatment [1]. The risk of recurrence is a significant concern for individuals who have undergone treatment for cancer. Various factors, i.e., the stage of initial cancer, specific biological markers (such as hormone receptor status), the

effectiveness of the initial treatment, and individual patient characteristics, are considered to influence the risk of recurrence [1]. With clinicopathologic characteristics of cancer patients, it is possible to predict 5-year cancer recurrence [5]. Doctors could use such a prediction to make a tailored treatment plan.

CAD systems leverage deep learning and multidisciplinary knowledge and techniques to analyze medical imaging and non-imaging data and provide the analyzed results to clinicians as second opinion or decision support in the various stages of the patient care process [6].

CAD tools, such as Aiforia, PathAI, Adjuvant!, PREDICT, and CanAssist-Breast (CAB) are among the popular ones. Aiforia [7] and PathAi [8] have similar capabilities, including automated image analysis, quantification of pathology features, and pattern recognition. Adjuvant! [9] and PREDICT [10] online tools are widely used in breast cancer recurrence prediction. Adjuvant! does not produce accurate results and is no longer available online [11]. PREDICT utilizes patient and tumor characteristics to generate predictions for individual patients. It helps clinicians and patients make informed decisions about treatment by estimating the likelihood of recurrence. Population-based study [12] conducted on older patients reported the effectiveness of PREDICT in 5-year recurrence and a slight overestimation in 10-year recurrence prediction. CAB [13] is another promising tool for the immunohistochemistry-based prognostic test; it utilizes biomarkers and clinical parameters, such as tumor size, grade and node status as inputs to generate a risk score and categorizes patients as low- or high-risk for distant recurrence within 5-years of diagnosis.

All the tools studied can perform only one histological image analysis or recurrence prognostics, not both. However, they have excellent 5-year recurrence results, while 10-year recurrence remains challenging.

The primary purpose of this research is to create a diagnostic CAD tool that can detect cancerous and non-cancerous areas in a breast cancer histological image, to predict cancer staging and to develop a generalized estimation of the risk of breast cancer 10-year recurrence, by combining the histological and clinical patient data. The paper describes in detail the methodology adopted for its development and the validation results. It is organized as follows: Section 2 shows background information and related work in machine learning and CAD tools; Section 3 describes the methodology adopted; Section 4 presents the experiments conducted and best results achieved; Section 5 discusses the experiment results and findings; and Section 6 concludes the paper by summarizing and providing future directions of work.

II. RELATED WORK

Machine learning has significantly advanced CAD tools in various ways, particularly in medical imaging, including breast cancer detection and diagnosis. CAD tools developed with conventional machine learning methods mainly use hand-engineered features based on the domain knowledge and expertise of human developers who translate the perceived image characteristics to descriptors that mathematical functions or conventional image processing techniques can implement. The recent advancement in computing power and dataset sizes allowed the application of deep convolutional neural networks (DCNN) to image classification problems. Contrary to the traditional approach of hand-crafted feature extraction methods, DCNNs learn useful features directly from the training image patches by optimizing the classification loss function.

Several studies focused on histological images with DCNNs. The works range from pioneering studies that introduced the concept of using deep learning for breast cancer diagnosis to sophisticated architectures tailored for specific tasks like segmentation [14] and feature extraction [15].

DCNN-based CAD systems can automatically extract meaningful features from histological images. These features include texture, color, shape, and intensity patterns. They can also perform image segmentation, which involves identifying and delineating specific regions of interest within the histological images. It is beneficial for isolating cancerous lesions or specific cell types. Usually, the cancer diagnosing process using a histological image consists of the following steps [16]. Firstly, tissue specimens are extracted through biopsy, affixed on glass slides, and stained with hematoxylin and eosin (H&E). Then, an expert histopathologist examines the glass slides under a light microscope to provide the diagnosis for each sample. Accurate interpretation of glass slides is crucial to avoid misdiagnoses, which require extensive time and effort by the pathologist. Each person could have up to a dozen biopsy samples that require analysis. It displays the necessity of computational digital pathology to augment and automate diagnosis processes by scanning digitized whole slide images (WSI). WSI contains many cells; the image could consist of tens of billions of pixels, which is usually hard to analyze. However, resizing the entire image to a smaller size, such as 256 X 256, would lead to the loss of information at the cellular level, resulting in a marked decrease in identification accuracy. Therefore, the entire WSI is commonly divided into partial regions of about 256 X 256 pixels (“patches”), and each patch is analyzed independently.

Araujo et al. [17] proposed a deep convolutional neural network combined with an SVM (support vector machine) to classify hematoxylin and eosin (H&E) stained histological images and achieved accuracies of 77.8% for four class (normal tissue, benign lesion, in situ carcinoma and invasive carcinoma) classification and 83.3% for two class (carcinoma, non-carcinoma) classification.

In [18], the popular DCNNs architectures pre-trained on ImageNet, such as VGG, ResNet and Inception extract the essential features from the images, and then gradient-boosted trees classifier is applied to classify the images.

Ensemble approaches like [15] and [19] also took similar approaches, except they employed an ensemble of DCNNs, namely VGG19, MobileNetV2 and DenseNet201, to extract visual features and then applied the boosting framework to achieve superior results in the detection of cancerous and non-cancerous areas from the histological image.

CAD tools, such as Aiforia and PathAI have cancerous and non-cancerous area detection features from a given histological image. Although breast cancer is detected early and the treatment is started soon after diagnosis, the cancer cells remain in the body undetected; after a certain period, it may recur. Machine learning methods are also applied to advance the prediction accuracy in breast cancer recurrence prediction. Usually, the datasets contain many features, which may mislead the prediction process as some features may lead to confusion or inaccurate prediction [20].

Feature selection is an essential first step in breast cancer recurrence prediction. In [20], a hybrid multi-stage learning technique based on brain-storming optimization was applied to study the most effective features, and it concluded that the feature selection is highly dependent on the applied classification algorithm and dataset used. [21] studied clinicopathologic characteristics of 579 breast cancer patients. It used statistical feature selection and particle swarm optimization to select and refine important features. It compared SVM, Decision Tree (DT) and Neural Network classifiers to predict breast cancer 5-year recurrence. It used the local invasion of the tumor, the number of tumors, the number of metastatic lymph nodes, the histological grade, the tumor size, estrogen receptor, and lympho-vascular invasion. PREDICT online tool predicts the recurrence based on features, such as breast cancer type, patient age, menopause, ER status, Ki-67, tumor size and tumor grade.

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 main contribution of this work is to develop an all-in-one CAD tool that utilizes machine learning algorithms like DCNNs and eXtreme Gradient Boosting (XGBoost). The developed CAD tool can generate a full breast cancer diagnostic and prognostic report by combining histological image analysis and clinical and histological data analysis.

III. METHODOLOGY

The development of the novel CAD tool consists of the following steps: data collection, dataset creation (stain normalization, patch extraction), model training and validation steps.

A. Data collection

Histological image analysis faces data variability, class imbalance, and potential bias challenges in general. Ensuring a representative and diverse dataset is crucial for training supervised DNN models that generalize well to real-world scenarios. It directly affects the performance of the trained model on new, unseen images. At the same time, accurate labels that indicate the presence or absence of the target condition, such as cancerous or non-cancerous tissue, are essential.

The data collection is achieved by digitizing the samples collected from anonymous patient biopsy slides provided by Verona Borgo Trento Hospital (Italy) with NED DP digital microscope. 300 sets of histological images with various magnification levels, specifically, 1.25x, 2x, 4x, 10x, 20x, and 40x, are collected and manually labelled by the Verona Borgo Trento hospital medical practitioners.

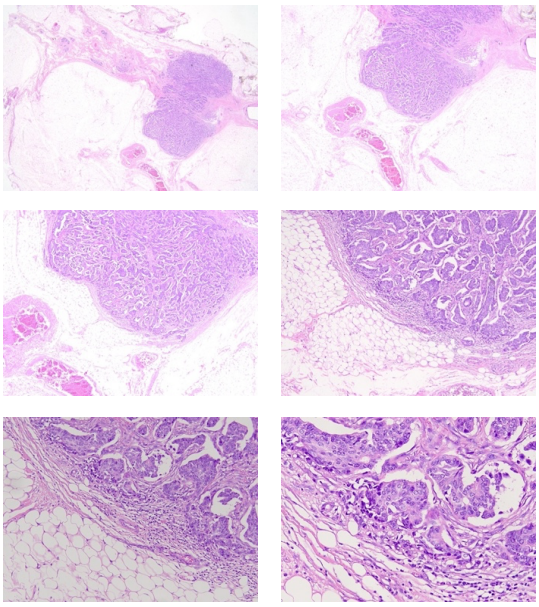

Figure 1. Different magnification variations (Starting from top left 1.25x, 2x, 4x, 10x, 20x, 40x)

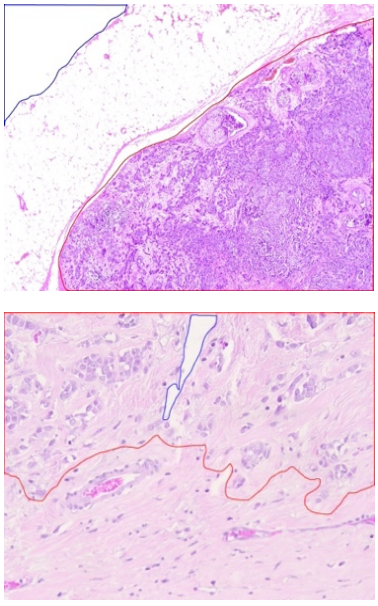

Figure 2. Labelled histological images

Each image has 1640x1175 resolution, and the labels are defined as blue and red color-coded lines on top of the tissue image, indicating areas without tissue cells and with cancerous tissue cells, respectively. As indicated in the following sample image, a closed-shape label makes it easy to separate and identify the areas in the next phase. Every image is also accompanied by clinical information and features, they are listed in the following table.

TABLE I. ALL FEATURES FROM CLINICAL DATA

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

B. *Dataset creation and model training*

The histological images are often stained to enhance the visibility of structures and cells. Variations in staining procedures can lead to differences in color and intensity, challenging comparing images or applying consistent analysis.

Firstly, considering the visual consistency and reproducibility of the experiments, stain normalization technique proposed in [22] is applied to the histological images to normalize the stains. Then, multiple datasets are created based on magnification levels with different image sizes. The clinical data was recorded by different medical specialists, each using a different structure and labelling. A preprocessing method standardizes the labels and filters out the missing data samples. The histological images without complete clinical data are discarded to prevent mismatching in the result. Secondly, the specific algorithms applied are chosen. The classification of cancerous and non-cancerous areas from the image is considered a binary classification, and the patch based DCNN approach is the best suited. The literature study shows that DCNN architectures like VGG and Inception pre-trained on ImageNet resulted in highly accurate classification models; therefore, fine-tuning the pre-trained models is favored.

A combination of multiple input sizes (patch dimension), various learning rates and batch sizes are experimented with, and the best accuracy model is selected at the end. Table II shows the specific parameters experimented during the model training. XGBoost is a popular machine learning algorithm known for its efficiency, speed, and performance in various predictive modelling tasks. After selecting the features with statistical feature selection, XGBoost is used in grade prediction.

TABLE II. VARIABLES EXPERIMENTED IN TRAINING

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

Finally, breast cancer recurrence is predicted with linear regression with the selected features. The experiment and result section describes the testing results regarding the parameters and model training.

IV. EXPERIMENTS

The experiment consists of preprocessing, image classification, grade prediction and recurrence prediction steps. In the experiments conducted, the images are pre-processed, multiple datasets are created by extracting multi-dimensional patches from them. Then, histological image classification, grade prediction and recurrence prediction models are trained. Combinations of various parameter are experimented, model accuracies and algorithms used are reported. The models are integrated to the CAD tool developed.

A. *Data preprocessing*

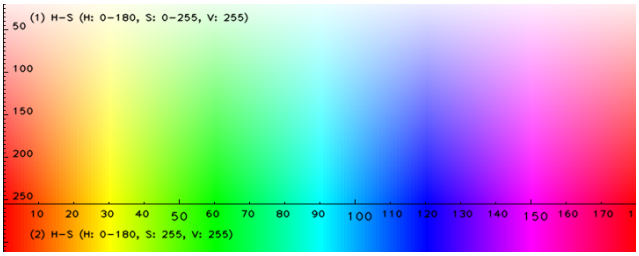

Figure 3. HSV color spectrum

In preprocessing, the color-coded image labels are separated with computer image processing techniques. With HSV color map spectrum (in Figure 3) and OpenCV, the blue and red color masks are created to separate the corresponding labels in the histological image. The preprocessing and construction of datasets are illustrated in Figure 4.

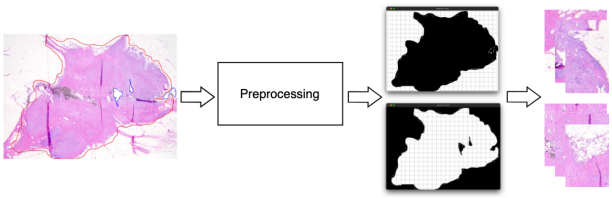

Figure 4. Preprocessing

Dilatation and erosion [23] techniques are also implemented to enhance the label continuity and fully surround the area of interest. Areas (in Figure 5) labelled with blue labels are excluded to minimize the impact of false classification. After successfully separating the color-coded labels, areas with negative and positive labels are represented with black and white masks. Then, patches are extracted from the corresponding areas to construct datasets.

A combination of 64x64, 150x150 and 200x200 patch sizes and magnification of 1.25x, 2x, 4x, 10x, 20x, and 40x are used to construct multiple datasets. The patch from the

edges contains positive and negative areas. A threshold of 0.7 is applied to label them. This threshold value is considered better suited because it produces relatively balanced datasets to work with. For instance, if 70% of the region is from a positively labelled region, it is labelled with a positive label.

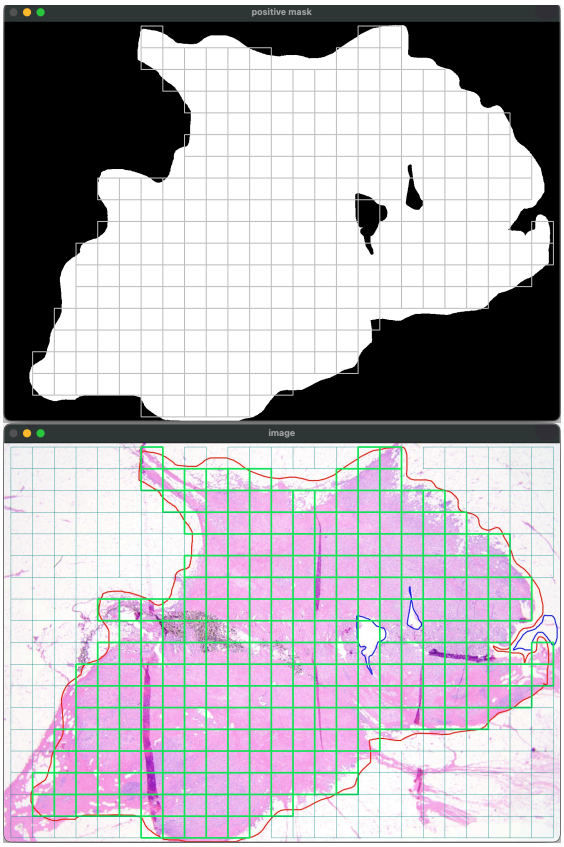

Figure 5. Label masks and patch extration

AWS Glue DataBrew [24] service is applied to establish a homogeneous dataset with corresponding clinical data and to facilitate the creation of an automated data pipeline, streamlining the data ingestion and preprocessing. Leveraging AWS SageMaker [25], essential data cleaning and preprocessing steps, such as feature scaling, categorical data encoding, and augmenting are conducted. To ensure the consistency of all the clinical data rows, "ISTOTIPO" is split into a set of derived variables namely, "ISTOTIPO_CDI", "ISTOTIPO_CLI", "ISTOTIPO_NST", "ISTOTIPO_TUBULARE", "ISTOTIPO_LOBULARE", "ISTOTIPO_APOCRINO", "ISTOTIPO_MICROPAPILLARE", "ISTOTIPO_MUCINOSO", "ISTOTIPO_PAPINCAPS" using one-hot vector encoding. This method provided a more comprehensive representation of the original "ISTOTIPO" feature and enriched the dataset, enhancing the model capacity for generating more accurate and insightful predictions. As a result of the preprocessing, a dataset of 300 data samples, each with 26 attributes, is constructed.

B. *Histological Image Classification*

The datasets constructed in the previous step are used to train DCNN models. Training, validation, and testing splits of 6:3:1 and 7:2:1 are experimented. Among all the trials conducted with different combinations of batch size, learning rate, input size (patch size) and DCNN architecture, fine-tuning pre-trained VGG16 on ImageNet with the following parameters (Table III) resulted in the best accuracy model.

TABLE III. THE BEST ACCURACY MODEL PARAMETERS

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

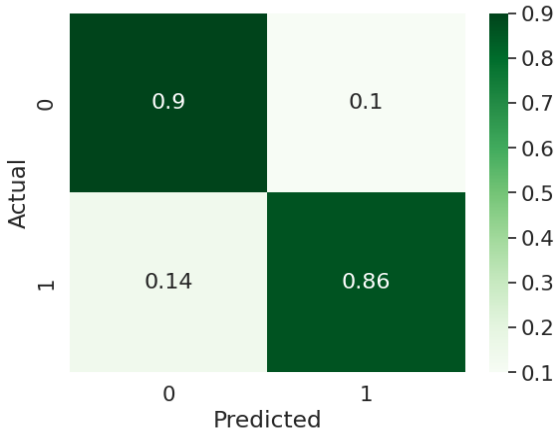

Figure 6. Label masks and patch extration

The confusion matrix of the model validation prediction is shown in Figure 6.

C. *Grade prediction*

The training process for this task revolved around a multiclass classification problem, where the goal was to categorize data into one of multiple predefined classes or categories. In multiclass classification, multiclass categorical cross entropy metric quantifies the differences between the predicted class probabilities and the true class labels for each data point. The features selected to train the model are listed in the Table IV below.

The model trained with XGBoost, achieved a validation accuracy of 95%. The multiclass logarithmic loss for the corresponding model is 0.21. The confusion matrix (in Figure 7) provides an in-depth insight into the model prediction, revealing its effectiveness in multiclass classification.

TABLE IV. ALL FEATURES FROM CLINICAL DATA

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

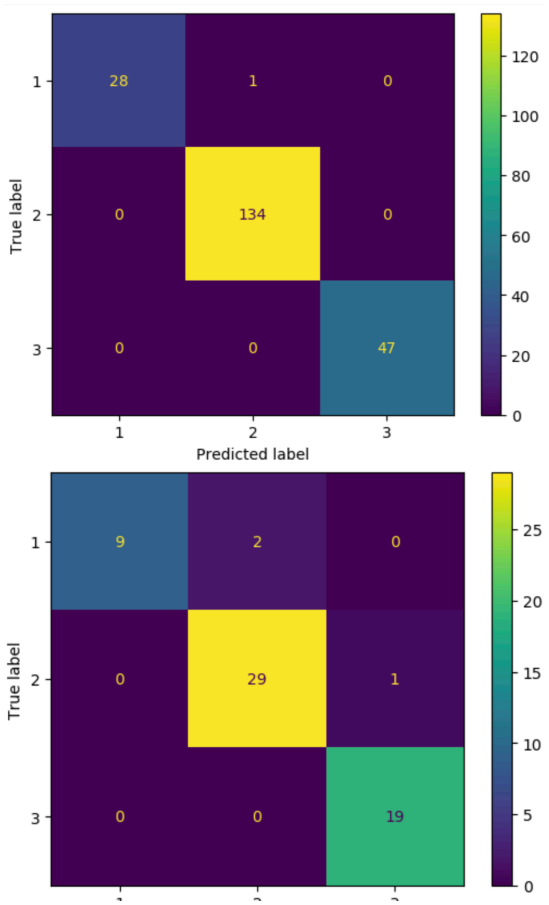

Figure 7. Label masks and patch extration

D. Breast cancer recurrence prediction

The correlations between features are explored and reduced to obtain a maximum accuracy model. However, 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%.

V. DISCUSSION

The accuracies achieved by image classification and linear regression are lower than to state-of-the-art results, especially with the linear regression. Fine tuning the pretrained VGG16 model achieved 87.6% accuracy, a reasonably good result but further investigations need to be conducted to improve the model robustness and accuracy. A comparison study should be conducted using open datasets to compare and validate the achieved classification result. The grade prediction with XGBoost algorithm achieved 95% accuracy, it demonstrates the effectiveness and efficiency of XGBoost algorithm. Finally, the linear regression for predicting the breast cancer 10-year recurrence only achieved 71% accuracy. Applying different machine learning algorithms, such as DT or SVM could improve the accuracy further. Regarding the data, greater magnification levels, such as 100x, 200x with more data certainly improve the over-all accuracies obtained in this study. Different subsets of features could be investigated in XGBoost and multiple linear regression to further improve the model accuracies. Moreover, different approaches, such as fusing multiple DCNNs as a feature extractor and combining different types of classifiers, such as SVM, DT or XGBoost could be the path to achieve a better result.

VI. CONCLUSION AND FUTURE WORK

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, and the cancer recurrence prediction with linear regression resulted 71% accuracy. It is an essential initial step in our future study direction. 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 studies will improve the accuracies and robustness of the models, acquire further labelled data and test with different DL approaches. Fusing molecular and genetic data and imaging feature might also enable a comprehensive understanding of disease characteristics.

REFERENCES

[1] 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.

[2] M. N. Gurcan et al. “Histopathological Image Analysis: A Review,” *IEEE Rev Biomed Eng*, vol. 2, pp. 147–171, 2009, doi: 10.1109/RBME.2009.2034865.

[3] J. G. Elmore *et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA - Journal of the American Medical Association*, vol. 313, pp. 10-11, 2015, doi: 10.1001/jama.2015.1405.

[4] 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.

[5] A. M. Gonzalez-Angulo *et al.*, "High risk of recurrence for patients with breast cancer who have human epidermal growth factor receptor 2-positive, node-negative tumors 1 cm or smaller," *Journal of Clinical Oncology*, vol. 27, pp. 33-34, 2009, doi: 10.1200/JCO.2009.23.2025.

[6] H. P. Chan, R. K. Samala, and L. M. Hadjiiski, "CAD and AI for breast cancer - Recent development and challenges," *British Journal of Radiology*, vol. 93, no. 1108. pp. 1-2, 2020. doi: 10.1259/bjr.20190580.

[7] "Aiforia." Accessed: 01.2024. [Online]. Available: https://www.aiforia.com/

[8] "PathAI." Accessed: 01.2024. [Online]. Available: https://www.pathai.com/

[9] P. M. Ravdin *et al.*, "Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer," *Journal of Clinical Oncology*, vol. 19, no. 4, 2001, doi: 10.1200/JCO.2001.19.4.980.

[10] "PREDICT." Accessed: 01.2024. [Online]. Available: https://breast.predict.nhs.uk/

[11] 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.

[12] N. A. De Glas *et al.*, "Validity of the online PREDICT tool in older patients with breast cancer: A population-based study," *Br J Cancer*, vol. 114, no. 4, pp. 1-2, 2016, doi: 10.1038/bjc.2015.466.

[13] M. M. Bakre *et al.*, "Clinical validation of an immunohistochemistry-based CanAssist-Breast test for distant recurrence prediction in hormone receptor-positive breast cancer patients," *Cancer Med*, vol. 8, no. 4, pp. 1755–1764, Apr. 2019, doi: 10.1002/cam4.2049.

[14] L. Yang, P. Meer, and D. J. Foran, "Unsupervised segmentation based on robust estimation and color active contour models," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 3, 2005, doi: 10.1109/TITB.2005.847515.

[15] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks," *CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, pp. 92–99, Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.11870

[16] M. Veta, J. P. W. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5. IEEE Computer Society, pp. 1400–1411, 2014. doi: 10.1109/TBME.2014.2303852.

[17] T. Araujo *et al.*, "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, Jun. 2017, doi: 10.1371/journal.pone.0177544.

[18] D. M. Vo, N. Q. Nguyen, and S. W. Lee, "Classification of breast cancer histology images using incremental boosting convolution networks," *Inf Sci (N Y)*, vol. 482, 2019, doi: 10.1016/j.ins.2018.12.089.

[19] K. Das, S. P. K. Karri, A. Guha Roy, J. Chatterjee, and D. Sheet, "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," in *Proceedings - International Symposium on Biomedical Imaging*, 2017. doi: 10.1109/ISBI.2017.7950690.

[20] M. Alwohaibi, M. Alzaqebah, N. M. Alotaibi, A. M. Alzahrani, and M. Zouch, "A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, 2022, doi: 10.1016/j.jksuci.2021.05.004.

[21] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," *Comput Struct Biotechnol J*, vol. 15, 2017, doi: 10.1016/j.csbj.2016.11.004.

[22] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," in *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, 2009. doi: 10.1109/ISBI.2009.5193250.

[23] "Eroding and Dilating." Accessed: 01.2024. [Online]. Available: https://docs.opencv.org/3.4/db/df6/tutorial_erosion_dilatation.html

[24] "AWS Glue DataBrew", Accessed: 01.2024. [Online]. Available: https://aws.amazon.com/glue/features/databrew/

[25] "AWS SageMaker", Accessed: 01.2024. [Online]. Available: https://aws.amazon.com/sagemaker/