PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY FERHAT ABBAS SETIF 1
FACULTY OF SCIENCES, DEPARTMENT OF PHYSICS





MASTER'S THESIS

Presented to obtain Master's degree in Medical Physics

PREDICTION OF BREAST CANCER RECURRENCE FROM HISTOPATHOLOGICAL IMAGES USING ARTIFICIAL INTELLIGENCE APPROACHES

Presented by: HEMADOU Manar

In front of the responsible jury:

CHOUABA SEIF E.	Pr. Université Sétif 1	Advisor
BARA Djemai	Dr. CDTA Algiers	President
BELKHIAT Djamel E.C.	Pr. Université Sétif 1 1	Examiner
KHOUDRI Saad	Dr. CLCC Sétif 1	Invited Member

18 JUNE 2026

Abstract

Breast cancer is the most frequently diagnosed cancer among women in Algeria, frequently identified at advanced clinical stages. Predicting the risk of disease recurrence is paramount for optimizing adjuvant therapeutic pathways and optimizing surveillance. This thesis introduces BCR-Net, a robust computational pipeline developed to explore recurrence prediction using weak supervision on digital histopathological images. The proposed architecture follows a two-stage paradigm: first, a sequential Convolutional Neural Network (CNN) Scorer filters and selects the top $K = 3000$ highly discriminative tissue patches from Whole-Slide Images (WSIs). Second, these patches are passed through a pre-trained ResNet50 model, spatially compressed using Global Average Pooling (GAP), and mapped into a compact 1024-dimensional feature space. A deep Multi-Instance Learning network with a Gated Attention mechanism (Att-MIL) is deployed to aggregate patch features and generate case-level outcomes. To benchmark the proposed framework, conventional machine learning classifiers including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forests (RF) were evaluated using extracted patch-level features . Furthermore, an interactive Graphical User Interface (GUI) was engineered using Tkinter to deploy and test the optimized architecture on a clinical cohort compiled in collaboration with the University Hospital Center (CHU) of Sétif. On the baseline test set, KNN reached a patch-level accuracy of 84.76% (F1-Score = 79.95%). When deployed on the local CHU Sétif dataset via the interactive interface, the Att-MIL model demonstrated high clinical sensitivity, achieving a patient-level recall of 54.10% and an overall classification accuracy of 71.40%. These results validate the capacity of automated computational pathology systems to serve as reliable prognostic tools within Algerian healthcare infrastructure.

Keywords

Breast cancer recurrence, Multiple Instance Learning (MIL), CNN discriminator, Digital pathology, Histopathology.

Résumé

Le cancer du sein est le cancer le plus fréquemment diagnostiqué chez les femmes en Algérie, souvent identifié à des stades cliniques avancés. Prédire le risque de récidive de la maladie est essentiel pour optimiser les parcours thérapeutiques adjuvants et la surveillance médicale. Cette thèse présente BCR-Net, un pipeline computationnel robuste développé pour explorer la prédiction de la récidive à partir d'images histopathologiques numériques sous supervision faible. L'architecture proposée repose sur un paradigme à deux étapes : premièrement, un classifieur basé sur un Réseau de Neurones Convolutifs (CNN) séquentiel filtre et sélectionne les K=3000 régions de tissu (patches) les plus discriminantes à partir d'images de lames entières (WSIs). Deuxièmement, ces patches sont traités par un modèle pré-entraîné ResNet50, compressés spatialement par Global Average Pooling (GAP), puis projetés dans un espace de caractéristiques de dimension 1024. Un réseau profond d'apprentissage par instances multiples avec un mécanisme d'attention guidée (Att-MIL) est déployé pour agréger ces caractéristiques et générer des prédictions à l'échelle du patient. Pour évaluer le pipeline, des approches classiques d'apprentissage automatique — K-Nearest Neighbors (KNN), Support Vector Machines (SVM) et Random Forest (RF) — ont été testées sur les caractéristiques des patches. De plus, une interface graphique interactive (GUI) a été développée avec Tkinter pour déployer et tester l'architecture sur une cohorte clinique réelle collectée en collaboration avec le Centre Hospitalier Universitaire (CHU) de Sétif. Sur le jeu de test de référence, le KNN a atteint une précision de 84,76% (F1-Score = 79,95%). Lors de son déploiement sur les données du CHU de Sétif via l'interface interactive, le modèle Att-MIL a démontré une sensibilité clinique élevée, atteignant un rappel de 54,10% et une précision globale de 71,40%. Ces résultats valident la capacité des systèmes de pathologie numérique automatisés à servir d'outils pronostiques fiables au sein des infrastructures de santé algériennes.

Mots-clés

Récidive du cancer du sein, Apprentissage par instances multiples (MIL), Discriminateur CNN, Pathologie numérique, Histopathologie.

خلاصة

سرطان الثدي هو السرطان الأكثر تشخيصا لدى النساء في الجزائر ، وغالبا ما يتم تحديده في مراحل سريرية متقدمة .يعد التنبؤ بخطر تكرار المرض أمرا ضروريا لتحسين الدورات العلاجية المساعدة والمراقبة الطبية .تقدم هذه الأطروحة بي سي آر-نت ، وهو خط أنابيب حسابي قوي تم تطويره لاستكشاف التنبؤ بالتكرار من الصور النسيجية الرقمية تحت إشراف ضعيف .ويستند الهيكل المقترح على نموذج من مرحلتين :أولا ، مصنف على أساس شبكة العصبية التلافيفية متتابعة)سي إن إن (مرشحات ويختار ك=3000 المناطق الأنسجة الأكثر تمييزا)بقع (من صور الشرائح بأكملها)القمة العالمية لمجتمع المعلومات .(ثانيا ، تتم معالجة هذه التصحيحات من قبل نموذج ما قبل إنترا إرمن ريسنيت 50 ، مضغوط مكانيا من قبل المتوسط العالمي تجميع)الفجوة (، ثم المتوقعة في الفضاء من خصائص البعد 1024 .يتم نشر شبكة التعلم متعددة المثيلات العميقة مع آلية الاهتمام الموجهة)أت-ميل (لتجميع هذه الخصائص وتوليد التنبؤات على مستوى المريض .لتقييم خط الأنابيب ، تم اختبار أساليب التعلم الآلي الكلاسيكية — ك-أقرب الجيران)كن (، وآلات ناقلات الدعم)سفم (والغابات العشوائية)رف (— على خصائص البقع .بالإضافة إلى ذلك ، تم تطوير واجهة رسومية تفاعلية مع تكينتر لنشر واختبار البنية على مجموعة سريرية حقيقية تم جمعها بالتعاون مع مركز المستشفى الجامعي)تشو (في سطيف .في مجموعة الاختبار المرجعي ، وصلت درجة الدقة إلى 84.76٪)درجة الفورمولا 1 = وخلال نشره على بيانات مستشفى سطيف عبر الواجهة التفاعلية ، أظهر نموذج العلاج الإشعاعي الموضعي .(79.95٪ حساسية سريرية عالية ، حيث وصل إلى استدعاء 54.10٪ ودقة إجمالية تبلغ 71.40٪ .تؤكد هذه النتائج قدرة أنظمة علم الأمراض الرقمية الآلية على العمل كأدوات تنبؤية موثوقة داخل البنى التحتية الصحية الجزائرية.

Contents

General Introduction 3

1 Medical Background 4
1.1 Introduction . . . 4
1.2 Background . . . 4
1.3 Breast Cancer . . . 4
1.4 Breast Cancer in Algeria . . . 5
1.5 Histopathological characteristics . . . 5
1.6 Methods of Obtaining Histopathology Images . . . 6
1.7 Breast Cancer Recurrence . . . 7
1.8 Clinical Importance of Recurrence Prediction . . . 7
1.9 Role of histopathology in breast cancer diagnosis . . . 8
1.10 Challenges in histopathological image analysis . . . 8
1.11 Oncotype DX (ODX) . . . 8
1.12 Conclusion . . . 9

2 CNN Architecture, ML Approaches 10
2.1 Introduction to Artificial Intelligence . . . 10
2.1.1 Machine Learning . . . 10
2.1.2 Deep Learning . . . 11
2.2 Artificial neural network . . . 11
2.3 Convolutional Neural Networks (CNNs) . . . 11
2.3.1 Introduction and Historical Development: . . . 11
2.3.2 Core Architecture and Components: . . . 12
2.3.3 Layers of CNN: . . . 12
2.4 Machine Learning Approaches: . . . 14
2.4.1 K-Nearest Neighbors (KNN) . . . 14
2.4.2 Support Vector Machine (SVM) . . . 14
2.4.3 Random Forest . . . 15
2.5 Optimization . . . 16
2.5.1 Adam Optimizer . . . 16
2.6 Related work . . . 16
2.7 Conclusion . . . 17

3 Materials and Methods 18
3.1 Introduction . . . 18
3.2 Software and Hardware Environment . . . 18
3.2.1 Hardware Configuration . . . 19

3.2.2 Software Tools and Libraries 19
3.3 Dataset Description 22
3.3.1 Public Dataset 22
3.3.2 University Hospital Center SETIF Dataset 22
3.4 Proposed Methodology 24
3.4.1 BCR-Net 24
3.5 Comparative Study 27
3.5.1 Machine Learning Models 28
3.6 Evaluation Metrics: 28
3.6.1 Confusion Matrix: 29
3.6.2 Accuracy: 29
3.6.3 Precision: 29
3.6.4 Recall (Sensitivity): 30
3.6.5 F1-Score: 30
3.6.6 ROC Curve and AUC: 30
3.7 Conclusion 30

4 Experimental Results and Discussion **31**
4.1 Introduction 31
4.2 Experimental Setup 31
4.2.1 Training and Test Configuration 31
4.2.2 Hyperparameter Settings 32
4.3 Results of CNN-Scorer 32
4.3.1 Training Performance Of CNN 33
4.4 Feature Extraction Results 33
4.5 Results of Multi-Instance Learning (MIL) 34
4.5.1 MIL Classification Results 34
4.5.2 Performance Evaluation 34
4.6 Comparative Study 35
4.6.1 KNN 35
4.6.2 SVM 36
4.6.3 Random Forest 37
4.6.4 Comparison with BCR-Net 38
4.7 Discussion 39
4.8 Test of UHC Data using BCR-Net 40
4.8.1 Inference and Clinical Testing Interface 40
4.8.2 Model Performance Metrics on CHU Sétif Dataset 42
4.9 Conclusion 42

Bibliography **45**