UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Deep Learning-Based Metastatic Detection and Localization for Breast Cancer Diagnosis in Whole Slide Histopathology Images

by

Sepideh Khaliliboroujeni

A Thesis Submitted
in Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2026

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Sepideh Khaliliboroujeni*, declare that this thesis is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *School of Electrical and Data Engineering, Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

Signature:

Production Note:
Signature removed prior to publication.

Sepideh Khaliliboroujeni

Date: 15 April 2026

ABSTRACT

Deep Learning-Based Metastatic Detection and Localization for Breast Cancer Diagnosis in Whole Slide Histopathology Images

by

Sepideh Khaliliboroujeni

Breast cancer is one of the most prevalent and lethal diseases among women worldwide, and early and accurate detection, particularly of metastasis, is crucial to improving patient survival rates. Deep learning techniques have revolutionized image analysis, yet their application to histopathology Whole Slide Images (WSIs) remains a challenge due to the massive size, complex tissue structures, and limited annotations of WSIs. Although advances such as patch-based classification, multi-scale modeling, and weakly supervised learning have improved accuracy and reduced annotation needs, challenges remain. Metastasis detection and localization approaches still face high computational demands, reliance on extensive annotations, and difficulty capturing global context and fine details. Patch-based methods are widely used as they enable manageable processing of large WSIs by dividing them into smaller regions. However, they often struggle with boundary precision and detecting small metastases, while reliance on high-resolution WSIs increases costs and limits scalability in clinical practice. This thesis addresses the core research question of how deep learning models can be optimized for metastasis detection and localization in large-scale histopathology images while ensuring efficiency, reliability, and clinical deployability in weakly supervised settings.

To address this question, this thesis proposes two novel, efficient, weakly supervised deep-learning approaches for the detection and localization of metastases in WSIs. The first approach, detailed in Chapter 3, focuses on maximal computational efficiency, achieving efficient inference for large-scale WSI analysis using

low-resolution inputs and minimal annotated data, making it suitable for rapid clinical triage.

The second, enhanced framework (Chapter 4) prioritizes spatial accuracy and improved data efficiency by addressing the limitations of coarse patch-based analysis. It improves contextual modeling and boundary-level precision through advanced architectural components, enabling more refined localization of metastatic regions under weak supervision. Both newly proposed methods reduce computational and annotation costs by requiring only a fraction of labeled data. Evaluated on the CAMELYON16 dataset, they outperform state-of-the-art methods and expert pathologists in metastasis detection and localization, demonstrating robustness in large-scale, noisy, and limited-data environments.

Dissertation directed by A/Prof. Wenjing Jia, Prof. Daniel Catchpoole and Prof. Massimo Piccardi.
School of Electrical and Data Engineering

Dedication

Dedicated to my husband, my love, my steadfast rock and unwavering support, whose love and belief in me have been my greatest strength.

To my sweet and beloved daughter, the brightest light in my life, the source of my fullest happiness, and the greatest inspiration in my heart.

To my dearest parents, whose unconditional love, sacrifices, and guidance have shaped who I am.

And to my lovely brothers, for their constant encouragement and support throughout this journey.

This work is a testament to the love and strength of my family.

Dedicated also to the pursuit of world peace and harmony.

Acknowledgements

I extend my heartfelt gratitude to my supervisor, Associate Professor Wenjing Jia, for her invaluable guidance, encouragement, expertise, and humanity, which have been instrumental in shaping this work. My deepest thanks also go to my co-supervisors, Professor Daniel Catchpoole, for his knowledge, generous support, and encouragement to explore new perspectives, and Professor Massimo Piccardi, for his gracious understanding, compassion, and support during challenging times.

I am profoundly grateful to my dearest husband, whose love and unwavering belief in me have been my greatest strength, and to my beloved daughter, whose unlimited love, joy and light have been my greatest inspiration throughout this journey. To my most precious parents, whose unconditional love and sacrifices have shaped who I am, and to my lovely brothers, whose constant encouragement and support have been a source of motivation—thank you for being my foundation and strength.

I also gratefully acknowledge the financial support from the Australian Government's Research Training Program Stipend Scholarship (RTPS), which made this work possible.

Finally, I would like to express my heartfelt gratitude to all those who have directly or indirectly contributed to my research. I extend special thanks to Professor Xiangjian He for his invaluable guidance and encouragement at the inception of my journey. I also sincerely thank Dr. Ye (Edward) Huang for his instrumental support during the early stages of my PhD and my co-authors for their valuable collaboration in enriching my publications.

Sepideh Khaliliboroujeni

Sydney, Australia, 2026.

List of Publications

Journal Article

1. **Sepideh Khaliliboroujeni**, Xiangjian He, Wenjing Jia, Saeed Amirgholipour. "End-to-end metastasis detection of breast cancer from histopathology whole slide images." *Computerized Medical Imaging and Graphics*, vol. 102, pp. 102-136, 2022. DOI: `https://doi.org/10.1016/j.compmedimag.2022.102136`.

Papers Under submission

1. **Sepideh Khaliliboroujeni**, Wenjing Jia, and Daniel Catchpoole, "Weakly Supervised Data-Efficient Framework for Metastasis Detection and Segmentation on Breast Cancer Histopathology Images." Under submission to *Computerized Medical Imaging and Graphics*, 2025.

Contents

Abstract iii

Dedication v

Acknowledgments vi

List of Publications vii

List of Figures xii

List of Tables xix

1 Introduction 1

2 Related Works 5

2.1 Introduction 5

2.2 Imaging Modalities for Breast Cancer Screening 9

2.2.1 Mammogram 9

2.2.2 Magnetic Resonance Imaging (MRI) 12

2.2.3 Ultrasound (US) 13

2.2.4 Histopathology 16

2.3 Deep Learning Approaches for Histopathology Image Classification . . 18

2.3.1 Multi-Scale and Resolution-Invariant Models 20

2.3.2 Hybrid and Ensemble Learning Architectures 23

2.3.3 Feature-based and Context-Aware Models 27

2.3.4 Adaptable Learning and Model Enhancement Strategies . . . 30

2.3.5 Foundation Models and Sequence Architectures in Histopathology . . . 37

2.4 Deep Learning Approaches for Metastasis Detection, Localization and Segmentation . . . 41

2.4.1 Metastasis Detection and Localization . . . 42

2.4.2 Metastasis Segmentation . . . 48

2.4.3 Multi-Task and Hybrid Models for Combined Detection, Localization and Segmentation . . . 54

2.5 Summary and Future Work . . . 67

3 An End-to-End Model for Metastasis Detection in Breast Histopathology Images 69

3.1 Introduction . . . 70

3.2 Comparison to Preceding Metastasis Detection Approaches . . . 73

3.3 The Proposed Method . . . 74

3.3.1 Network Architecture . . . 75

3.3.2 Loss Functions . . . 78

3.3.3 Training and Inference . . . 80

3.4 Experiments . . . 83

3.4.1 Dataset . . . 83

3.4.2 Evaluation Metrics . . . 84

3.4.3 Qualitative Evaluation . . . 85

3.4.4 Ablation Studies . . . 85

3.4.5 Quantitative Evaluation and Discussion . . . 88

3.4.6 Computation Time . . . 91

3.4.7 Implementation Details . . . 92

3.4.8 Discussion and Limitation 93

3.5 Summary 94

4 A Weakly-Supervised Data-Efficient Framework for Metastasis Detection and Localization in Breast Histopathology Images 97

4.1 Introduction 99

4.2 Related Work 102

4.2.1 Feature Ensemble Methods 102

4.2.2 Attention Mechanisms in Histopathology Image Analysis . . . 104

4.2.3 Multi-scale Analysis 106

4.2.4 Weakly Supervised Learning 108

4.2.5 Motivations 109

4.3 The Architecture of the Proposed Model 110

4.3.1 Input Preprocessing and Initial Feature Extraction 114

4.3.2 Integration of GRUs with ResNet 115

4.3.3 Bottleneck Feature Refinement 119

4.3.4 CBAM in WSI Processing 120

4.3.5 Dual-Task Framework and Loss Computation 123

4.3.6 Decoder Architecture for High-Resolution Metastasis Localization 125

4.4 Implementation and Evaluation 130

4.4.1 Dataset 134

4.4.2 Evaluation Metrics 134

4.5 Experiment Results 134

4.5.1 Visual Evaluation 134
4.5.2 Objective Evaluation 138
4.5.3 Sensitivity Analysis of Lesion Scoring Strategies 140
4.5.4 Controlled Evaluation and Robustness Analysis 141
4.5.5 Boundary-Level Evaluation 144
4.6 Ablation Studies 145
4.6.1 The Effectiveness of the Bi-GRUs in the Encoder 146
4.6.2 The Effectiveness of ASPP in the Bottleneck 147
4.6.3 The Effectiveness of the CBAM [173] and skip connections [132] 148
4.6.4 The Effectiveness of loss weightings on overall model performance 150
4.7 Performance Efficiency Evaluation 151
4.7.1 External Validation on an Independent Dataset 152
4.8 Summary and Discussion 153

5 Summary and Conclusion 156

List of Figures

2.1 Mammogram Modality: There is cancer in the right breast, but as can be seen, it is difficult to be recognized because it is a very tiny lesion [1]. 10

2.2 Illustration of the DBT. While the detector is immobile or spins slightly, the breast is pressed and fixed, and then the X-ray roll turns at several angles towards that. *Source: [59].* 11

2.3 (A) A full-field digital mammogram (FFDM) of a patient with a lesion. (B) DBT of the same patient presents an extremely clear lesion compared to FFDM. *Source: [89].* 12

2.4 Examples of MRI. *Source: [1]* 13

2.5 Examples of normal, benign, and malignant ultrasound images. *Source: [5].* 14

2.6 B-mode or brightness mode ultrasound (the classical process of obtaining an ultrasound scan) (a) and ultrasound Nakagami imaging of a benign breast lesion (b). *Source: [158].* 14

2.7 A benign tumor and malignant tumor of the breast by SWE. *Source: [3].* 15

2.8 Comparison of different imaging technologies in spotting a tumor in a breast. *Source: [1, 115].* 15

2.9 A whole slide imaging of a Stage 4 lesion in a biopsy sample. Patches surrounded by red squares present Stage 4 lesions' objective attributes that have been distinguished visibly in a high-resolution schema; Patches enclosed by blue squares have not been distinguished as they only represent attributes of lower-stage lesions. *Source: [117].* 17

2.10 Histopathological imaging samples of different subgroups of breast cancer from BreaKHis [147, 123] dataset. *Source: [68].* 18

2.11 Left: a sample histopathology whole slide image. Right: A patch that has been magnified in high resolution is shown. *Source: [75].* . . 18

2.12 Overview of the ADSVM framework [190]. The pipeline consists of three main stages: (1) **Pre-processing**, where images undergo normalization and subdivision into patches; (2) **Patch screening**, which uses local feature descriptors, locality-constrained linear coding, and an SVM classifier to identify and correct mislabeled patches; and (3) **RANet classification**, which performs final classification using a resolution-adaptive network. *Source: [190].* . . . 21

2.13 The DEEP_Pachi model block diagram: (a) feature extraction using ensemble backbone models (DenseNet201 and VGG16), (b) patch embedding, (c) position embedding, (d) DEEP_Pachi components—self-attention network and MLP layer, and (e) testing stage with new images on the trained DEEP_Pachi network. *Source: [159].* 24

2.14 The UMTL framework. The figure illustrates TPLG's pseudo-label generation, TLC's label refinement, and mutual learning for WSI classification. *Source: [74].* 26

2.15 Overview of the DeepSMILE workflow: illustrating the contrastive self-supervised pre-training, variability-aware MIL process, and final classification pipeline. *Source: [137].* 31

2.16 A two-stage architecture of the method proposed in [30], where the first model (Nc) quickly identifies mitosis candidates, and the second model (Nf) refines this by accurately distinguishing mitoses from similar structures, improving both speed and accuracy [30]. . . . 43

2.17 An overview of the two main components of Her2Net: the convolutional section and the deconvolutional section [134]. 48

2.18 The SegMitos model [93] detects mitosis using concentric labeling, marking the mitotic area and excluding surrounding regions from loss calculation. 49

2.19 Tile-wise segmentation approaches explored in the study [126]: (a) direct tile-wise merging of segmentation results, and (b) tile-wise merging using a Conditional Random Field (CRF) for improved accuracy. 50

2.20 A detailed overview of the PartMitosis system [140]. The weights of the prediction layer in the strong segmentation branch (W_{Strong}) are derived from the weights in the concentric segmentation branch (W_{Concent}) through a weight transfer function (Γ), which is not learned as a model parameter. 56

2.21 The DDTNet architecture [186], integrating detection and segmentation modules with a feature fusion strategy for precise lymphocyte segmentation. 56

2.22 The framework of SA-MIL [96], which uses VGGNet stages and self-attention modules for feature extraction, along with deep supervision and a fusion layer to integrate multi-scale predictions. . . 58

2.23 The overall workflow of the models [76]: (a-b) Fresh breast tumor tissues stained and imaged; (c-d) High-resolution images captured using a microscope; (e-g) Low-resolution images simulated and processed by SRGAN-ResNeXt model; (h-i) High-resolution outputs generated; (j-k) Inception U-net Model segments the images. 58

2.24 An illustration of the mitosis detection method proposed in [72]. (a) data preprocessing, and (b) a two-stage system, 'Detecting Fast' for rapid candidate segmentation and 'Detecting Slow' for detailed refinement of candidates. 60

2.25 The BreasTDLUSeg framework, comprising two main components: (A) SPPC-Net for identifying TDLU-positive patches and (B) PPS-Net for precise pixel-level segmentation within these patches. . . 61

3.1 Pyramid representation of a whole slide image, where the level-0 slide (bottom) represents the principal WSI with the highest resolution and the level-n slide presents the smallest image by the base size divided by 2^n [117]. 70

3.2 Examples of histopathological image patches exhibiting sporadic noise. 71

3.3 Examples of healthy (the first row) and unhealthy (the bottom row) histopathological image patches. The diversity of tissue structure and sample appearance is observable, highlighting the challenge of distinguishing between normal and tumor patches due to their visual similarities. 71

3.4 An overview of the model. 75

3.5 The framework of the model for training (a) and prediction (b). . . . 76

3.6 An example of applying the Otsu method on a WSI. The non-tissue background is filtered using the Otsu thresholding method [121]. . . . 80

3.7 An example of a predicted mask before (left) and after (right) applying the morphological operations on the predicted mask. The green boxes surround some tiny detected points that are hardly visible. The zoomed-in views of the cross-sectional areas show that some small detected spots are removed after applying the morphology operations. 83

3.8 (a) The left column shows the whole slide images where tumor polygons are drawn over them (these contours are indicated in blue), (b) the middle column shows the ground truth of tumor regions in the WSIs, and (c) the third column shows the output of the method. The white boxes show some small tumor areas that are not easy to detect. This method still detects them explicitly. The red boxes show some tricky, healthy spots that also are identified accurately. . . 87

3.9 Visualization of the breast cancer detected with the proposed method. (a) The whole slide image is overlaid with tumor polygons (indicated in blue). (b) The ground truth of tumor regions. (c) The predicted tumor regions by the proposed method. 94

3.10 (a) The ground truth mask along with the zoomed-in view of a small area of the tumor mask, and (b) the predicted mask along with the corresponding zoomed-in view of the same area. A slight difference can be observed between the ground truth and the detected mask. . . 94

4.1 Overview of the proposed architecture for metastasis detection and localization in histopathology images. Input patches are processed by a ResNet-50 + Bi-GRU encoder. Multi-scale contextual features are refined at the bottleneck using ASPP and CBAM. The decoder reconstructs dense localization maps via transposed convolutions with CBAM-enhanced skip connections. Dual-task learning is performed through classification and tumor ratio regression at both the bottleneck (global features) and decoder (high-resolution features). The outputs from the bottleneck and decoder branches are combined via averaging to produce the final classification and regression predictions, enabling joint metastasis classification and spatially resolved localization. 111

4.2 The detailed framework of the proposed model. 113

4.3 An example of a WSI (left) and several extracted patches from it (right). 114

4.4 Illustration of Bi-GRUs embedded in ResNet for feature refinement. Feature maps from each ResNet layer are reshaped into sequences and processed by Bi-GRUs to capture spatial dependencies. 117

4.5 Illustration of the CBAM module integrated into the proposed framework. The top section shows the **Channel Attention Module**, which refines the input feature map by emphasizing informative feature channels. The bottom section depicts the **Spatial Attention Module**, which further enhances the intermediate features by focusing on important spatial regions. The input feature map is progressively refined to produce the final output feature map, improving metastasis localization. 121

4.6 Detailed illustration of a single decoder stage in the proposed architecture. 129

4.7 Comparison between ground truth annotations and the predicted binary metastasis masks generated by the proposed method. 135

4.8 Detailed analysis of a selected region of interest (ROI) within a WSI. The first column shows the input WSI, with the selected region highlighted by a green border. The second column presents the corresponding ground truth mask at higher resolution, while the third column shows the binary mask predicted by the proposed method. 136

4.9 A comparison of tumor localization results across different methods. (a) WSI with metastasis regions outlined. (b) Hi and Li [63], (c) Wang *et al.* [169], and (d) Guo *et al.* [57] show probability maps, where warmer colors indicate higher tumor likelihood. (e) Probability map generated by the proposed method, visualized using a red–blue color scheme for consistent comparison, and (f) ground truth annotation. 137

4.10 Figure 4.10: Comparison of localization results under different model configurations. (a) Original WSI with ground truth metastasis boundaries (blue) and selected region of interest (ROI) highlighted in green. (b) Zoomed-in ROI with ground truth annotation. (c) Probability map generated by the proposed model. (d)–(f) Probability maps obtained from ablated variants of the model without Bi-GRUs, ASPP, and CBAM with skip connections, respectively, where warmer colors indicate higher tumor likelihood. . . 147

List of Tables

2.1 Overview of 2016-2024 papers on BC classification in HP images using deep learning. 34

2.2 Comparison with modern WSI architectures 40

2.3 Overview of 2016-2024 papers on BC detection, localization and segmentation in HP images using deep learning. 63

3.1 Details of datasets, the numbers of images used for training and testing in the CAMELYON16 dataset (Bejnordi textitet al.) [21]. . . 84

3.2 Comparison of the results with and without specific modules. 88

3.3 Comparison of results with top-ranking methods on CAMELYON16 [21]. For the proposed method, WSI-level split results are reported as mean ± standard deviation using WSI-level k-fold cross-validation ($k = 5$). 90

3.4 Comparison of the processing time of the proposed method with previously reported methods for processing an ROI and a WSI. . . . 92

4.1 Comparison with representative prior methods on CAMELYON16, including the proposed model at 10× using 40 WSIs.(official test set) 139

4.2 Sensitivity analysis of lesion-level scoring strategies for FROC evaluation on the CAMELYON16 dataset. 141

4.3 Controlled comparison on CAMELYON16 (official test set) at 10× using 40 WSIs with stain normalization. Primary results use WSI-level k-fold cross-validation (mean ± std across folds); the patch-split setting is reported as a single-run reference. 143

4.4 Performance of the proposed method using the full CAMELYON16 training set (270 WSIs) at 10× with stain normalization. 144

4.5 Boundary-level overlap performance on the CAMELYON16 test set at 10× with stain normalization. WSI-level results correspond to the primary evaluation protocol of this chapter; the patch-split row is provided as a single-run reference. 145

4.6 Ablation results obtained by removing individual modules while keeping the training and evaluation protocol fixed under the WSI-level setting with stain normalization. 146

4.7 Ablation study of classification and regression loss weights $(\lambda_{\text{class}}, \lambda_{\text{reg}})$ under the WSI-level protocol with stain normalization. Results are reported using AUC (WSI-level detection), FROC (lesion localization), and boundary-level Dice/IoU. Bold indicates the best-performing configuration. 150

4.8 Comparison of processing time for the proposed method and representative existing methods for processing an ROI and a full WSI under comparable inference settings. 151

4.9 Comparison of in-domain performance on CAMELYON16 and external validation on CAMELYON17. The CAMELYON17 results reflect generalization performance obtained without re-training or fine-tuning under the same inference pipeline. 152

List of Abbreviations

- ADC - Analog-to-Digital Converter
- ASPP - Atrous Spatial Pyramid Pooling
- AUC - Area Under the Curve
- BCE - Binary Cross Entropy
- CAD - Computer-Aided Diagnosis
- CAM - Class Activation Map
- CAMELYON - Cancer Metastases in Lymph Nodes (Challenge Dataset)
- CBAM - Convolutional Block Attention Module
- Chl_Atn - Channel Attention module
- CLAHE - Contrast Limited Adaptive Histogram Equalization
- Clf - Classification
- CNN - Convolutional Neural Network
- DBN - Deep Belief Network
- DBT - Digital Breast Tomosynthesis
- DCIS - Ductal Carcinoma In Situ
- Fref - CBAM refined feature maps from encoder skip connections
- FC - Fully Connected
- FCN - Fully Convolutional Network
- FFDM - Full-Field Digital Mammography
- FROC - Free-Response Receiver Operating Characteristic

- GAP - Global Average Pooling
- GPU - Graphics Processing Unit
- GRU - Gated Recurrent Unit
- H - Height
- H&E - Hematoxylin and Eosin
- HRD - Homologous Recombination Deficiency
- IDC - Invasive Ductal Carcinoma
- IG - Information Gain
- MIB - Magnification-Independent Binary
- MIL - Multiple Instance Learning
- MRI - Magnetic Resonance Imaging
- MSB - Magnification-Specific Binary
- MSE - Mean Squared Error
- PCA - Principal Component Analysis
- RANet - Resolution Adaptive Network
- Reg - Regression
- ROI - Region of Interest
- SFM - Screen-Film Mammography
- SIE - Single Instance Evaluation
- SOP - Second-Order Pooling
- SoS2MIL - Second-Order Self-supervised Multiple Instance Learning
- Spt_Atn - Spatial Attention

- SWE - Shear Wave Elastography
- TC - Transposed Convolution
- TCGA - The Cancer Genome Atlas
- TLC - Transformer Label Cleaner
- TPLG - Transformer Pseudo-Label Generator
- U (Ui, Uup) - Decoder feature maps at stage i and upsampled
- US - Ultrasound
- VGG - Visual Geometry Group
- W - Width
- WSI - Whole Slide Image

Chapter 1

Introduction

Breast cancer is a leading cause of illness and death among women globally, representing approximately 30% of all newly diagnosed cancer cases. This disease arises from genetic mutations that disrupt normal cell growth regulation, resulting in uncontrolled cell proliferation. The most prevalent form, invasive ductal carcinoma, accounts for nearly 80% of diagnoses and is characterized by its potential to spread beyond the initial tumor site, a process known as metastasis. Detecting metastatic spread early is essential for determining treatment options, improving patient prognosis, and enhancing survival rates.

The field of breast cancer diagnostics has been transformed by advancements in imaging technology and machine learning. Traditional imaging techniques such as X-ray mammography, MRI, and ultrasound have proven valuable in identifying breast abnormalities, but histopathological imaging remains the benchmark for detailed cellular-level analysis. Histopathology involves the microscopic examination of tissue samples stained with hematoxylin and eosin (H&E), which reveals structural changes indicative of cancer. The digitization of these samples into whole slide images (WSIs) allows for comprehensive, high-resolution analysis and remote pathology assessments.

However, the analysis of WSIs presents significant challenges due to their large file size, which can reach gigabytes per image. While high-resolution WSIs provide unparalleled detail, they also demand extensive computational resources and complex algorithms for effective analysis. Furthermore, manual annotation of WSIs is

an exhaustive process requiring considerable expertise, limiting the volume of data available for training deep learning models.

In recent years, various deep learning-based technologies have revolutionized image analysis by enabling automated feature extraction, pattern recognition, and decision-making with high accuracy. Unlike traditional computer vision techniques that rely on handcrafted features, deep learning models—particularly Convolutional Neural Networks (CNNs)—learn hierarchical representations directly from raw image data. These models excel in tasks such as classification, detection, localization, and segmentation, where they identify objects, localize regions of interest, and assign pixel-wise labels, respectively.

The success of deep learning in natural image processing has extended to specialized domains, including medical image analysis, where accurate and efficient interpretation is crucial for clinical decision-making. In medical imaging, deep learning is applied across various modalities, such as MRI, CT scans, X-rays, and histopathology images, to detect diseases, classify abnormalities, and segment pathological regions. Deep learning has significantly improved computer-aided diagnosis (CAD) systems, enhancing accuracy and reducing the burden of manual pathology assessment. However, challenges such as data scarcity, high computational demands, and the need for generalizable models continue to drive advancements in this field.

Existing solutions have demonstrated significant improvements in diagnostic accuracy and consistency, reducing the subjectivity associated with human evaluation. Nonetheless, the unique challenges posed by WSIs, such as high computational demands and limited annotated data, necessitate innovative approaches to ensure robust and scalable CAD systems. Many current models incur high computational costs, making them less feasible for deployment in computationally constrained settings. Additionally, there is a heavy dependency on annotated datasets, which can

be a significant bottleneck due to the scarcity and expense of acquiring expertly labeled medical images. A crucial limitation is the inability to effectively capture contextual relationships within the images. This issue primarily stems from traditional patch-based processing methods, which focus on localized areas and thus miss the integral global context necessary for accurate localization. This often results in suboptimal performance, especially in scenarios requiring a detailed understanding of complex anatomical structures. Furthermore, these models exhibit narrow adaptability, often performing well only within the confines of the specific conditions and data types for which they were trained. This restriction in versatility becomes apparent when models are applied to new or varied datasets, highlighting the need for more robust models with enhanced adaptive capabilities. Additionally, boundary refinement remains a substantial challenge. Many models struggle to accurately delineate complex anatomical structures, resulting in less precise localization particularly at object boundaries, which could adversely affect subsequent medical analysis and treatment planning.

The motivations of this research are to balance accuracy and efficiency, ensuring fast and reliable results while reducing dependence on costly expert annotations. A key objective is to enhance the model's understanding of the context within breast tissue structures and improve accuracy at tumor boundaries for precise diagnostics. Additionally, the research focuses on optimizing the model's attention to critical tumor-specific features. Finally, the aim is to design advanced deep learning technologies that can support clinical workflows, enabling practical and effective computational pathology analysis.

This thesis is structured as follows: Chapter 2 provides a comprehensive literature review of deep learning applications for the analysis of breast cancer in histopathology images. The review covers classification models as well as detection, localization, and segmentation approaches, and further discusses recent ad-

vances in foundation models and sequence-based architectures for whole-slide image analysis. Chapter 3 and 4 detail two novel end-to-end models for metastatic detection and localization in WSIs. Chapter 3 presents a framework that prioritizes computational efficiency and rapid inference for large-scale WSI analysis. In contrast, Chapter 4 introduces an enhanced framework that prioritizes spatial precision and boundary-level localization by incorporating advanced context-aware and multi-scale modeling components. Specifically, Chapter 3 presents an approach integrating attention mechanisms for efficient classification and localization, leveraging lower-resolution data to improve processing speed and reduce annotation requirements. Chapter 4 presents an advanced model integrating Gated Recurrent Units (GRUs) [33], Atrous Spatial Pyramid Pooling (ASPP) [31], Convolutional Block Attention Module (CBAM) [173], and skip connections [132] within a residual recurrent attention framework. This architecture enhances the spatial accuracy of metastasis localization and boundary-level precision by capturing long-range dependencies, multi-scale context, and fine spatial details while focusing on tumor-relevant features. Finally, Chapter 5 summarizes the findings, discusses their practical implications, and outlines future research directions aimed at further improving deep learning methodologies for breast cancer diagnostics.

Chapter 2

Related Works

The emergence and evolution of deep learning techniques have significantly advanced automated image processing techniques for predicting breast cancer development. Recent progress in deep learning-based classification, detection, localization, and segmentation of breast cancer in histopathological images has substantially improved automated diagnostic capabilities. This chapter provides a comprehensive review of high-impact deep learning methods in this domain, systematically examining each study's objectives, methodologies, datasets, key findings, contributions, and limitations. The review begins with an overview of different medical imaging modalities, highlighting how histopathology (HP) images offer the most detailed phenotypic information for monitoring cancer metastasis. The review then focuses on two critical aspects of breast cancer diagnosis: 1) histopathology image classification, and 2) metastasis detection, localization, and segmentation. The review concludes by identifying current research gaps and suggesting future directions to advance the field.

2.1 Introduction

Breast cancer develops due to genetic mutations that disrupt normal cell growth regulation. These genetic changes lead to unchecked cell division and proliferation, resulting in the formation of cancer [62, 149, 165]. Breast cancer typically originates in the milk-producing lobules or the ducts that carry milk to the nipple but can also develop in the fatty or fibrous connective tissues within the breast. Approximately 1 in 20 women worldwide will be diagnosed with breast cancer during their lifetime,

with projections indicating that by 2050, new cases and deaths will increase by 38% and 68%, respectively, with a disproportionate burden expected in countries with a low Human Development Index [84]. Early detection is critical as it enables more effective treatments, reduces healthcare costs, and significantly improves survival rates.

Over the years, radiologists and pathologists have employed various imaging techniques, including X-ray mammography, magnetic resonance imaging (MRI), ultrasound, histopathologic imaging, and digital breast tomosynthesis (DBT), to detect and monitor breast cancer. While these imaging modalities are valuable, they also have limitations. For instance, overlapping breast tissues can obscure suspicious masses, complicating early detection [112]. Histopathology images, derived from biopsy samples stained with hematoxylin and eosin (H&E), offer a much deeper level of detail. Unlike other imaging modalities, which provide a macroscopic view, histopathology allows the examination of cellular and subcellular structures, revealing subtle morphological changes indicative of cancerous growths. This detailed view aids in distinguishing between different cancer subtypes, such as benign versus invasive ductal carcinoma, which is crucial for determining treatment strategies.

With advancements in digital scanning, whole slide images (WSIs) have emerged as a powerful tool in cancer diagnosis, allowing for high-resolution views of tissue samples across multiple magnifications. Each WSI can be as large as one gigabyte, representing tissue samples at different resolution levels in a pyramidal structure. The highest resolution (Level-0) provides a detailed view, while subsequent levels progressively reduce resolution, facilitating efficient data access and analysis. This multi-resolution format enables pathologists to explore both detailed and broad perspectives, supporting an accurate diagnosis and making histopathology images more adaptable to automated analysis. Digital histopathology also allows for electronic storage, transmission, and analysis through deep learning models, offering higher

diagnostic accuracy and consistency than traditional imaging.

This literature review focuses on the applications of deep learning, particularly convolutional neural networks (CNNs), for breast cancer diagnosis using histopathology images. Traditionally, histopathological image analysis has relied on manual assessment by pathologists, which is time-consuming and can be subject to variability. Deep learning offers new possibilities for automating and enhancing the accuracy of breast cancer classification, detection, localization, and segmentation in histopathology images. Classification, detection, localization, and segmentation each play critical roles in computational histopathology, particularly in the context of metastasis diagnosis.

Classification, especially at the patch and WSI levels, is widely used to distinguish between healthy and abnormal tissue and to provide slide-level predictions that guide downstream analysis. In weakly supervised settings, WSI classification plays a key role in identifying whether a slide contains metastatic regions, serving as a high-level diagnostic decision that supports downstream spatial analysis. It enables efficient slide triaging and reduces reliance on dense annotations by guiding attention to slides most likely to contain lesions. Classification methods form the basis for many WSI-based detection and localization strategies. Patch-level classification is frequently utilized to support WSI-level classification, often determining whether a slide contains metastases—an important step in detection workflows. Furthermore, classification outcomes are widely integrated into detection and segmentation models, particularly in weakly supervised settings, where they guide the localization of metastatic regions.

Detection refers to identifying the presence and approximate location of suspicious regions, serving as a key bridge between high-level classification and low-level spatial analysis. Localization builds on this by refining the position and extent

of metastases, providing more actionable insight for diagnosis and treatment planning. Segmentation, on the other hand, enables fine-grained, pixel-level delineation of tumor boundaries, offering critical spatial information that enhances diagnostic precision. In the context of metastasis, accurate segmentation contributes directly to localization by refining the spatial extent and morphology of suspicious regions, especially in cases involving small or irregularly shaped lesions. Together, these techniques form the foundation for robust and interpretable computer-aided diagnosis systems in histopathology.

This review is organized into the following sections:

- **Section 2.2** provides an overview of major imaging techniques for breast cancer detection.
- **Section 2.3** reviews studies focused on breast cancer classification using deep learning on histopathology images, including recent advances in foundation models and sequence-based architectures for whole-slide image analysis.
- **Section 2.4** explores advancements in breast cancer detection and segmentation with deep neural networks on histopathology images.
- **Section 2.5** concludes the review by summarizing key insights and identifying open research challenges for future work.

By synthesizing the significant contributions in this field, this review highlights the critical role of deep learning in advancing breast cancer diagnosis, offering enhanced precision, efficiency, and scalability in the analysis of histopathological images.

2.2 Imaging Modalities for Breast Cancer Screening

There are several methods for breast screening in detecting abnormalities and cancer, such as mammography, MRI, ultrasound, and histopathologic imaging. In the following, each method is reviewed one by one.

2.2.1 Mammogram

A mammogram (MG) is an X-ray image of breast tissue that can help investigate breast cancer in people who have no clear symptoms. There are three types of mammogram images *i.e.*, screen-film mammography (SFM), full-field digital mammograms (FFDM), and DBT.

Screen-film Mammography (SFM) is a traditional method for diagnosing cancer in the early stage and has supreme sensitivity in diagnosing fatty breast tissue [40, 146, 44]. It applies X-ray radiation to generate images of the breast, which have been used for breast cancer classification in different studies [37].

However, although SFM is an acceptable modality in detecting fatty tissue, it has low sensitivity in detecting dense breast tissue containing fatty tissue, milk glands, milk ducts and supportive tissue. These tissues look more solid and white on a mammogram, so it is not easy to spot them using SFM (Fig. 2.1). 10-20% of dense breast tissue cannot be identified by this method [146, 44, 25].

Digital Mammography (DM), also known as full-field digital mammography (FFDM), is an imaging modality that uses X-ray. But instead of X-ray film in SFM, electronics is used to convert X-rays into mammographic images of the breast. This modality has been an effective criterion model in detecting abnormality in early-stage breast cancer [37, 73]. The other advantages of this modality include progress in images for possible better confronting and improved sensitivity compared with SFM [44, 122].



Figure 2.1 : Mammogram Modality: There is cancer in the right breast, but as can be seen, it is difficult to be recognized because it is a very tiny lesion [1].

However, there are some disadvantages to this technology as well. First, DM has a high false-positive rate (low specificity). So, people without cancer may test positive and therefore potentially have to go through an anxiety pathway and also receive unnecessary diagnostic procedures [37, 73]. In addition, DM incurs high costs compared with SFM [73] and high radiation exposure [136]. Also, it is more likely to fail to detect cancer for overlapping breast tissue [53, 80].

Digital Breast Tomosynthesis (DBT) is a new image modality and is also called 3D mammography. It can help both radiologists and CAD systems to enhance their reputation by detecting breast cancer more precisely [37, 58, 52, 6, 168, 125, 184]. This model applies low-extent X-rays by moving the pipe with the X-ray in a circular arc around the breast. It then constructs a three-dimensional picture from the received information at several angles of the breast [67]. An illustration of this method is shown in Fig. 2.2.

There are several worthy advantages of this modality. DBT enhances the ability to diagnose all types of breast cancer abnormalities compared to 2D digital mam-

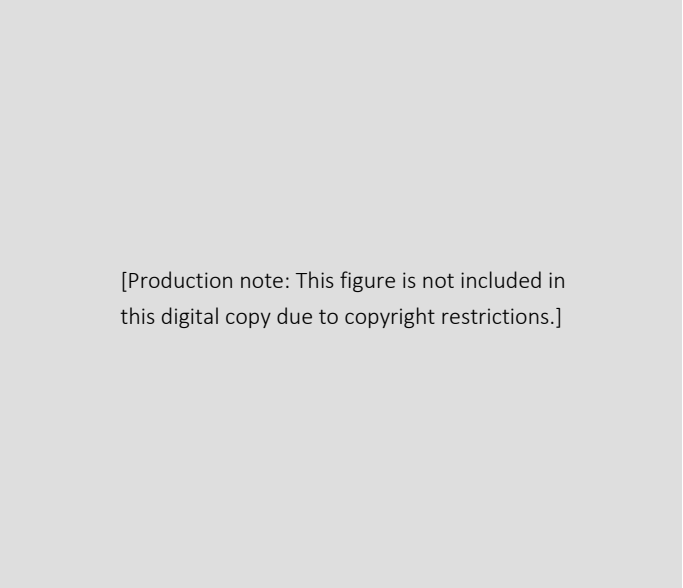


Figure 2.2 : Illustration of the DBT. While the detector is immobile or spins slightly, the breast is pressed and fixed, and then the X-ray roll turns at several angles towards that. *Source: [59].*

mography [67]. Fig. 2.3 shows a comparison of the clarity of an FFDM and a DBT of a patient's lesion [89]. Furthermore, the difficulty of detecting overlapping tissue has been leveraged by this modality. In fact, DBT technology's quiddity removes overlapping breast tissue. The DBT technology leads to breaking up and visualizing overlapping breast tissue in separate fragments, so the image is recreated without overlapping on that [37, 90, 131, 13, 64, 83]. Also, DBT improves the reliability and precision of imaging in terms of detecting cancer and can achieve better specificity, sensitivity, and accuracy [88, 129] and lower false-negative and false-positive rates compared to 2D DM [37, 136, 103, 34, 128].

Meanwhile, there are some disadvantages to this imaging process as well. For instance, the amount of radiation in DBT is increased compared to FFDM, which side effects need to be considered [51, 124]. Moreover, taking DBT is more costly compared to other modalities [109].

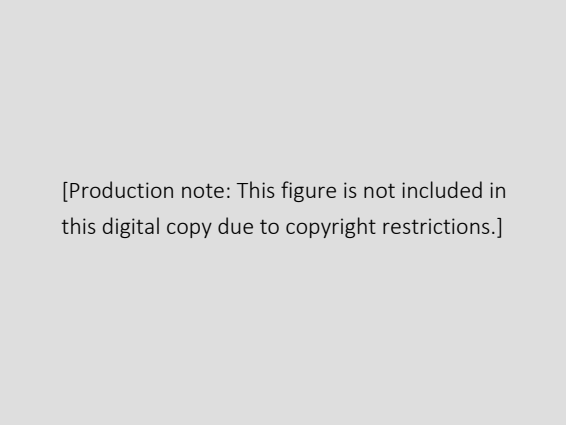


Figure 2.3 : (A) A full-field digital mammogram (FFDM) of a patient with a lesion. (B) DBT of the same patient presents an extremely clear lesion compared to FFDM. *Source: [89].*

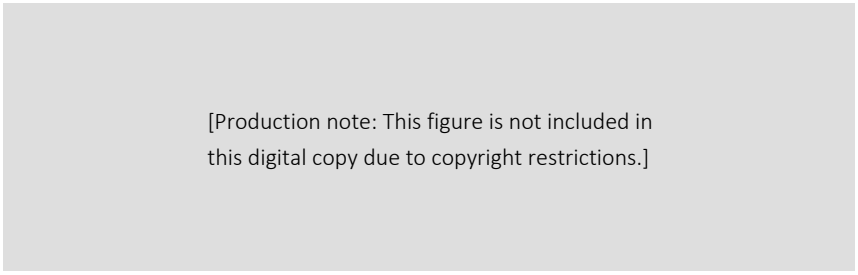
2.2.2 Magnetic Resonance Imaging (MRI)

MRI is an imaging modality, which instead of using radiation, applies magnetic flux and radio waves to create a more comprehensive image of the internal parts of the body. A visualization of the application of MRI in detecting tumors can be seen in Fig. 2.4, where in Fig. 2.4(a), a progressive tumor has been sighted, and in Fig. 2.4(b), an early-stage tumor is visible.

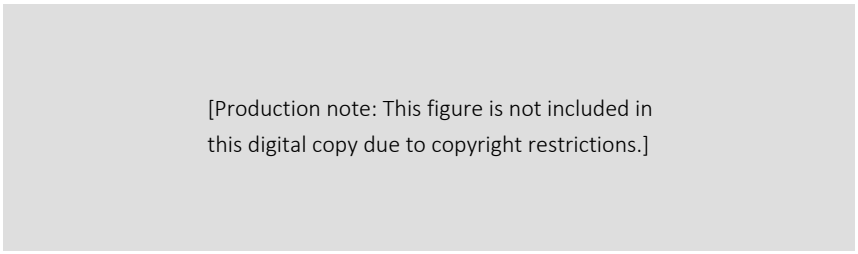
This methodology helps to recognize tumors and different irregularities where X-rays are weak or unable to detect them [154].

Fig. 2.8 shows that using MRI gives a clearer vision of a tumor [1, 115] compared to other technologies. Also, MRI provides a better result of sensitivity compared to ultrasound and digital mammogram methodology [115, 100, 7]. Besides, MRI has the capability called MRI-guided biopsy, which is the ability to distinguish regions in a breast that can be applied for biopsy when it is required [154, 115].

However, there are some unfavorable facts about MRI that should be considered as well. For example, MRI is too expensive. Additionally, MRI is not practical for



(a) A progressive tumor in the left breast.



(b) An early-stage tumor in the left breast.

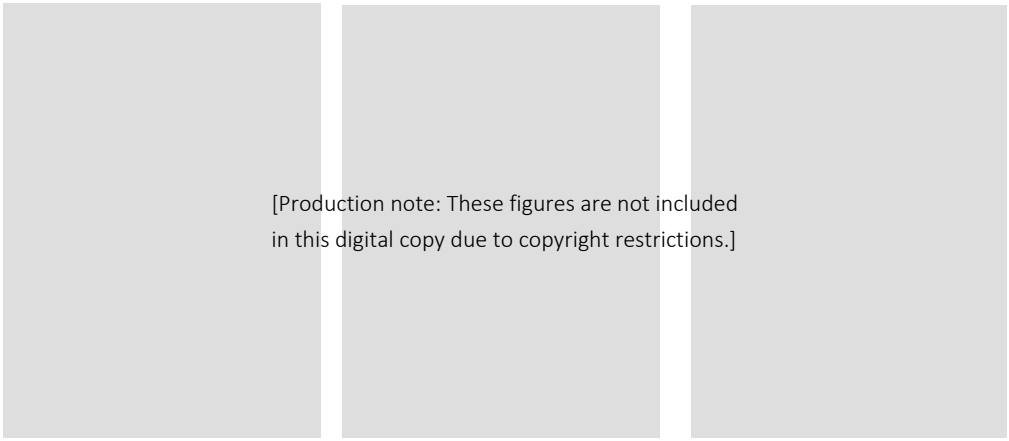
Figure 2.4 : Examples of MRI. *Source: [1]*

discerning malignant glands and benign glands in some cases, *e.g.*, in the detection of breast fibroadenomas, which contributes to most false-positive outcomes. Also, for doing MRI, the patient will be surrounded by a closed space, which may cause some difficulty for people, especially for those who have claustrophobia, mental illness [37, 55, 23].

2.2.3 Ultrasound (US)

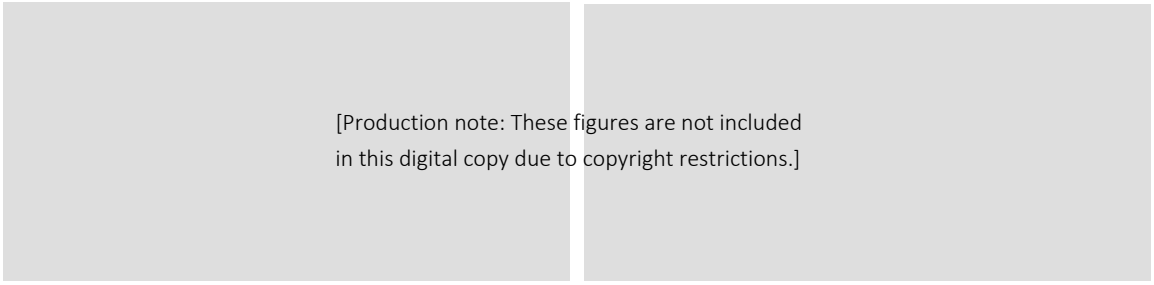
Ultrasound modality, also called sonography or, in the case of capturing the heart, echocardiography, applies high-frequency sound waves to produce live images of organs inside the body. Fig. 2.5 shows examples of normal, benign, and malignant breast images.

Furthermore, advanced ultrasound, such as ultrasound with Nakagami images (see Fig. 2.6) and shear-wave elastography (SWE) (see Fig. 2.7), has been created, providing superior attributes.



(a) Normal (b) Benign (c) Malignant

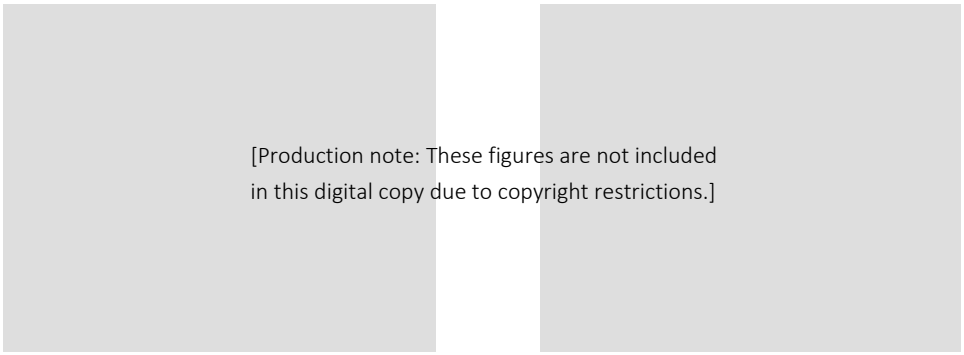
Figure 2.5 : Examples of normal, benign, and malignant ultrasound images. *Source: [5].*



(a) (b)

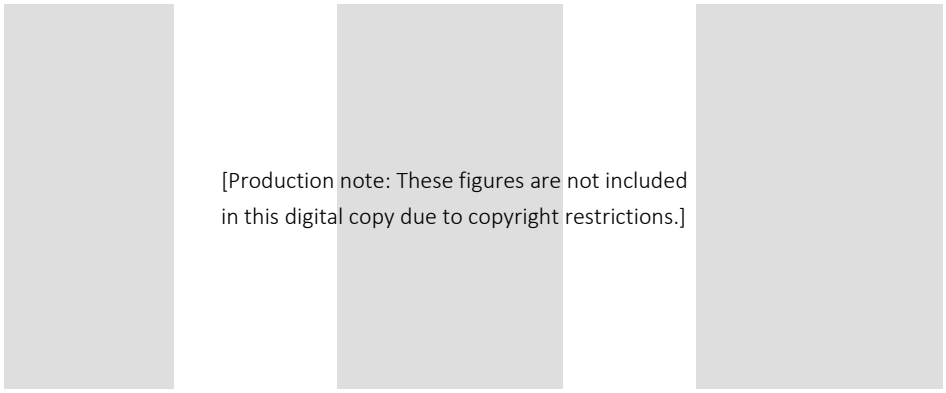
Figure 2.6 : B-mode or brightness mode ultrasound (the classical process of obtaining an ultrasound scan) (a) and ultrasound Nakagami imaging of a benign breast lesion (b). *Source: [158].*

In comparison to DM, MRI, and DBT, ultrasound is a popular, precise imaging method, which is not expensive and does not have any hazardous effects [143, 26]. Also, the US detects lesions in dense breasts with a good amount of recall [151, 17]. Moreover, there is an image-guided process called ultrasound-guided (US-guided), which helps recognize invasive cancer places to be applied later for biopsy [115].



(a) Benign (b) Malignant

Figure 2.7 : A benign tumor and malignant tumor of the breast by SWE. *Source: [3].*



(a) Mammogram (b) Ultrasound (c) MRI

Figure 2.8 : Comparison of different imaging technologies in spotting a tumor in a breast. *Source: [1, 115].*

Meanwhile, the US is a dependent technology, as imaging is impossible without an operator's assistance [26]. Also, the analysis of ultrasound images is complex [37, 143, 26]. Furthermore, calcifications that are symptoms of cancer in a primitive stage cannot be diagnosed by using ultrasound. Consequently, only using ultrasound for detecting cancer is not a very reliable methodology, and in most conditions, other breast imaging technologies such as MRI and mammograms need to be done along [54, 5].

2.2.4 Histopathology

The inspecting process of detecting, analyzing, locating, and classifying an invasive lesion or another anomaly biopsy fragment under a microscope is called histopathology. For that, biopsy samples extracted from the abnormal areas of the breast are established on glass sheets. In the following, the amount of contradiction in histopathology biopsy will be cut down by applying stain normalization preprocessing on these glass sheets [115]. Furthermore, the stained sheets are transformed into WSI, which present corresponding colored digital pictures of the biopsy samples. In breast cancer, the stage refers to how far the cancer has spread. Stage 4 is the most advanced stage, where cancer has spread to distant parts of the body. Fig. 2.9 [117] shows an example of a whole slide scan of a biopsy containing a Stage 4 lesion. Red squares highlight areas that show features commonly seen in advanced cancer. Blue squares indicate regions that look more like lower-stage or normal tissue. This shows how cancer features can vary across different parts of the same slide.

For detecting different cancer subgroups, as shown in Fig. 2.10 [68], regions of interest (ROI) patches are commonly drawn from WSI by specialist pathologists in several magnification aspects. Fig. 2.11 [75] shows a sample histopathology WSI image and its sample patch extracted from the image.

Among all imaging modalities, histopathology is often regarded as the "gold standard" [36]. However, there are several advantages and also some restrictions for histopathology imaging. For instance, images can be kept and sent electronically. Also, applying histopathology imaging compared to other scanning methods leads to more desirable outcomes in analyzing cancer. Additionally, the possibility of creating several ROIs from WSI reduces the chance of inability to detect cancer, and accordingly, the false-negative rate is decreased as well [115].

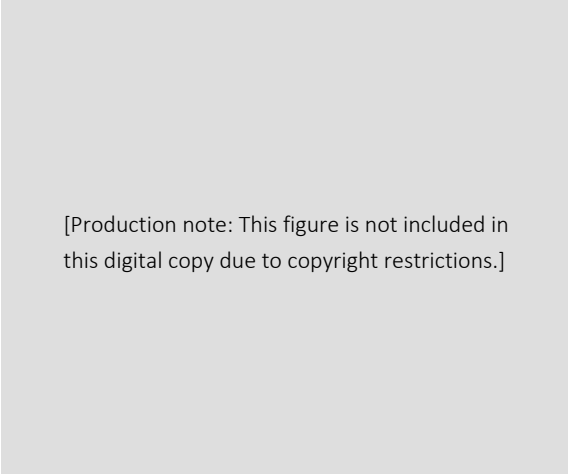


Figure 2.9 : A whole slide imaging of a Stage 4 lesion in a biopsy sample. Patches surrounded by red squares present Stage 4 lesions' objective attributes that have been distinguished visibly in a high-resolution schema; Patches enclosed by blue squares have not been distinguished as they only represent attributes of lower-stage lesions. *Source: [117].*

Alongside, there are also a few disadvantages for HP imaging, including outward changes of histopathology images through multiple elements, such as the variety of lab contracts, the difference in pathologists' expertise level, color conversion, and diversity in preprocessing by staining images, which lead to misdiagnosis [115, 110].

Furthermore, when using HP images as data, complicated techniques, a large amount of memory, and advanced computational equipment are necessary to enhance artificial intelligence diagnostics. In regard to that, classical machine learning methods applying HP images provide poor outcomes because an identical subgroup of cancer may appear very differently across images [115, 110].

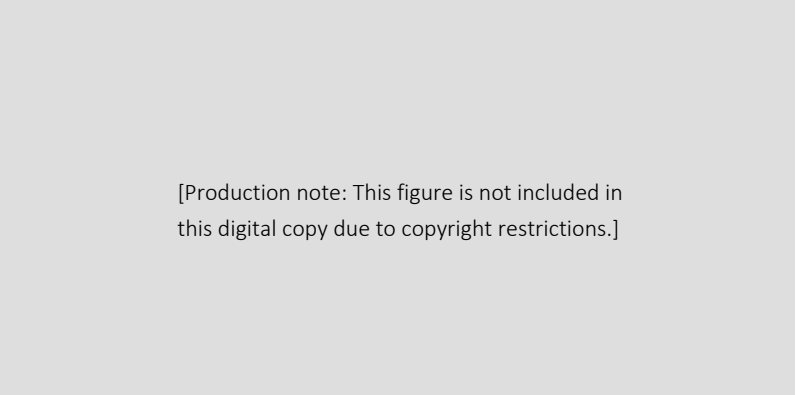


Figure 2.10 : Histopathological imaging samples of different subgroups of breast cancer from BreaKHis [147, 123] dataset. *Source: [68].*

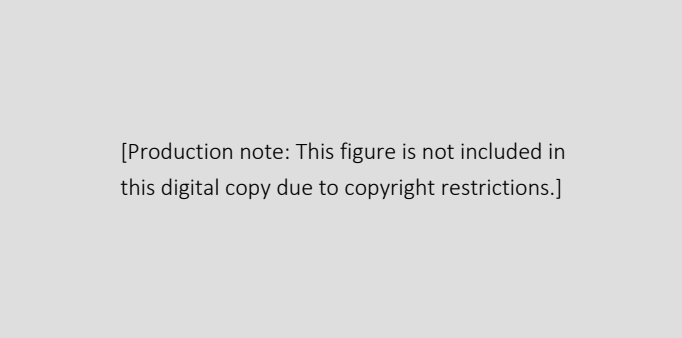


Figure 2.11 : Left: a sample histopathology whole slide image. Right: A patch that has been magnified in high resolution is shown. *Source: [75].*

2.3 Deep Learning Approaches for Histopathology Image Classification

Histopathology image classification, particularly at the patch level, serves as a critical foundation for several downstream tasks in computational pathology. It enables region-level scoring and heatmap generation, which are widely used to guide metastasis localization and pixel-wise segmentation, especially in weakly supervised settings where dense annotations are limited. Patch-based classification not only contributes to fine-grained diagnostic insight but also plays a central role in produc-

ing attention maps for region proposal and lesion highlighting. When aggregated, these patch predictions support WSI-level classification, which is essential for triaging slides and determining whether metastasis is present at the slide level. WSI-level classification supports effective slide triaging and is increasingly used in clinical pipelines to prioritize slides for further analysis. In this work, WSI classification is performed alongside localization, enabling the model to simultaneously assess slide-level diagnostic relevance while identifying the spatial distribution of metastatic regions. Reviewing classification methods is therefore key to understanding the foundations upon which many detection, localization, and segmentation strategies are built. This section reviews both binary classification models, which distinguish between benign and malignant cases, and multi-class classification models, which further differentiate between various breast cancer subtypes. Some studies also report patient-level classification performance, providing broader clinical relevance.

Additionally, our insights frequently consider how classification methods may facilitate downstream tasks such as detection, localization, and segmentation, acknowledging their foundational role in enabling more advanced diagnostic frameworks. Key research studies on classifying histopathology images for breast cancer diagnosis are categorized into four large groups based on the approaches they mainly used to provide a cohesive view of the advancements in breast cancer classification. These four subsections are: *Multi-Scale and Resolution-Invariant Models*, *Hybrid and Ensemble Learning Architectures*, *Feature-Based and Context-Aware Models*, and *Adaptable Learning and Model Enhancement Strategies*. The categorization adopted in this section is based on the *primary methodological contribution* emphasized by each study, rather than on a strict or mutually exclusive taxonomy. In histopathology image analysis, modern deep learning frameworks frequently integrate multiple techniques—such as multi-scale processing, attention mechanisms, transfer learning, and weak supervision—within a single pipeline. Consequently,

overlap between methodological dimensions is expected and widely acknowledged in the literature [101, 86]. To ensure clarity and analytical consistency, each study is assigned to the category that reflects its dominant architectural or learning strategy. This categorization scheme follows established practices in computational pathology reviews and facilitates systematic comparison of methods and trends [101, 28].

2.3.1 Multi-Scale and Resolution-Invariant Models

This category focuses on deep learning models designed to handle histopathology images captured at varying resolutions and magnifications. These architectures incorporate scale-adaptive designs and multi-resolution processing strategies to ensure consistent accuracy across varying scales and resolutions, enabling reliable cancer detection and progression assessment across diverse imaging setups. By emphasizing robustness to magnification changes and resolution variability, these models support scalable deployment across different slide scanners and imaging protocols.

Bayramoglu *et al.* [16] introduced a CNN-based approach addressing magnification independence in histopathology images. They proposed single-task and multi-task architectures. Using the BreaKHis [147] dataset, the single-task CNN achieved 83.25% accuracy, while the multi-task CNN reached 82.13% accuracy in malignancy classification and 80.10% in magnification estimation. The method enables robust performance across multiple scales without retraining, providing an early foundation for scale-agnostic and magnification-invariant modeling. However, its generalization in complex diagnostic tasks is limited by uneven data distribution across magnifications.

Building on scale-aware design, Zhou *et al.* [190] proposed a hybrid method combining anomaly detection (ADSVM) and a Resolution Adaptive Network (RANet) to further optimize classification across magnifications. As illustrated in Fig. 2.12, the ADSVM framework includes pre-processing, patch screening, and RANet clas-

sification. This approach achieved a high computational efficiency and 98.83% and 99.14% accuracy on the BreaKHis [147] and BACH 2018 datasets [9], respectively at 200× magnification. However, performance decreased at 400× magnification, and separate handling of multi-resolution feature maps remains a limitation.

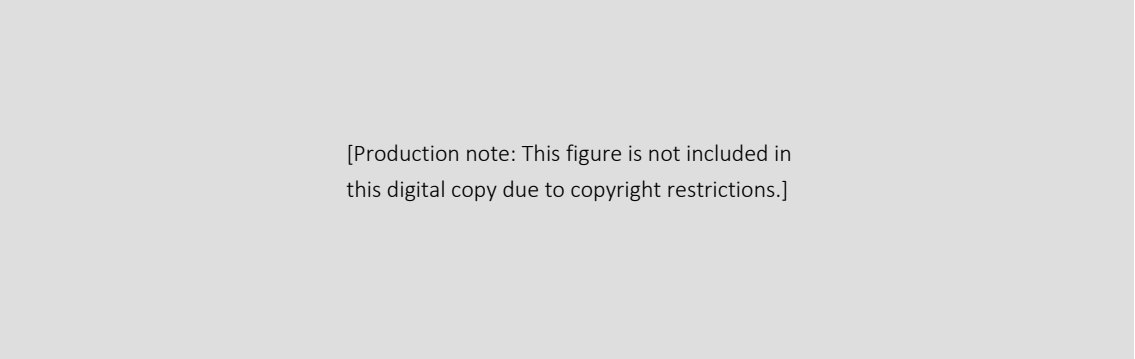


Figure 2.12 : Overview of the ADSVM framework [190]. The pipeline consists of three main stages: (1) **Pre-processing**, where images undergo normalization and subdivision into patches; (2) **Patch screening**, which uses local feature descriptors, locality-constrained linear coding, and an SVM classifier to identify and correct mislabeled patches; and (3) **RANet classification**, which performs final classification using a resolution-adaptive network. *Source: [190].*

Extending multi-scale learning, Taheri and Golrizkhatami [153] extended classification strategies with DenseNet201-based systems. They introduced magnification-specific binary (MSB) and magnification-independent binary (MIB) classifiers. On the BreaKHis dataset [147], the MIB classifier achieved a patient-level accuracy (PLA) of 93.67%, indicating the model's robustness across different magnification levels. However, it faced challenges with inter-patient variability, and score-level fusion may limit interpretability. Similarly focusing on magnification robustness, Xiao *et al.* [175] leveraged Inceptionv3 with transfer learning to classify breast cancer images efficiently across magnification levels. Using an image partitioning technique,

the model achieved over 92% accuracy on BreaKHis [147]. While effective in managing input size constraints, broader validation is still needed. Furthermore, reliance on partitioning without attention-based fusion may limit the capture of cross-scale dependencies.

Collectively, these studies emphasize the significance of magnification independence and scale flexibility, introducing complementary approaches to advance diagnostic robustness in breast cancer classification. Bayramoglu *et al.* [16] laid a strong foundation by demonstrating that CNNs can generalize across magnifications, though generalization remained limited in complex cases. while Zhou *et al.* [190] advanced this with resolution-adaptive networks. Taheri and Golrizkhatami [153] addressed magnification variation by combining predictions from multiple magnification levels, contributing an adaptable model. Xiao *et al.* [175] furthered these efforts by segmenting high-resolution images while maintaining classification accuracy and improving efficiency. However, cross-scale dependency modeling remained limited.

While each methodology offers valuable contributions, key challenges persist. These include limited performance at higher magnifications, reduced generalization across patients, and insufficient integration of multi-resolution features. The reviewed models primarily rely on decision-level fusion or partitioned inference strategies, which restrict their ability to capture spatial continuity and contextual relationships across scales. Magnification-independent models enhance diagnostic workflows with scalable solutions. However, they face limitations in handling patient variability and high-magnification performance. Overall, although they achieve strong classification performance, their limited ability to model inter-patch relationships and spatial dependencies constrains their effectiveness in representing complex tumor structures in whole slide images.

2.3.2 Hybrid and Ensemble Learning Architectures

This category focuses on classification models that integrate multiple architectures or learning modules—such as CNNs with traditional classifiers, RNNs, or transformers—to enhance diagnostic performance through architectural diversity and modular design. These hybrid and ensemble frameworks improve generalization, robustness, and adaptability across varying breast cancer subtypes and imaging contexts.

Araújo *et al.* [8] proposed a CNN-based CAD system combined with SVM for classifying biopsy images, showing improved accuracy at image-level classification However, the lack of deeper feature refinement modules or attention mechanisms limits its ability to focus on diagnostically relevant regions adaptively.

Zhu *et al.* [191] further enhanced hybrid learning by proposing an ensemble of compact dual-branch CNNs structure with Squeeze-Excitation-Pruning (SEP) blocks balancing local and global features while reducing overfitting. Nevertheless, aggressive pruning compromised spatial detail retention. Both approaches highlight limitations in capturing complex structures and boundary precision without advanced refinement modules like attention or recurrent mechanisms.

To address spatial dependencies, Yan *et al.* [179] introduced a hybrid model, combining Inception-V3 and bidirectional LSTM for 4-class classification. While this approach effectively combines residual networks for feature extraction and RNNs for spatial modeling, the reliance on LSTM networks introduces inefficiencies due to their higher computational complexity and susceptibility to vanishing gradient issues. Similarly, Srikantamurthy *et al.* [148] developed a CNN-LSTM hybrid model, CNNs for spatial feature extraction and LSTMs for sequential dependencies, to classify benign and malignant breast cancer subtypes. This approach effectively captures sequential context across patches but introduces substantial computational

overhead and faces scalability challenges across magnification levels.

Incorporating attention for spatial refinement, Ukwuoma *et al.* [159] proposed DEEP_Pachi, an ensemble-based hybrid model, combining DenseNet201 and VGG16 with self-attention mechanisms to capture global and regional features. A detailed illustration of the model is presented in Fig.2.13 [159]. Tested on BreaKHis [147] and ICIAR 2018 datasets, it achieved strong accuracy but raised concerns of potential overfitting and limited validation diversity.

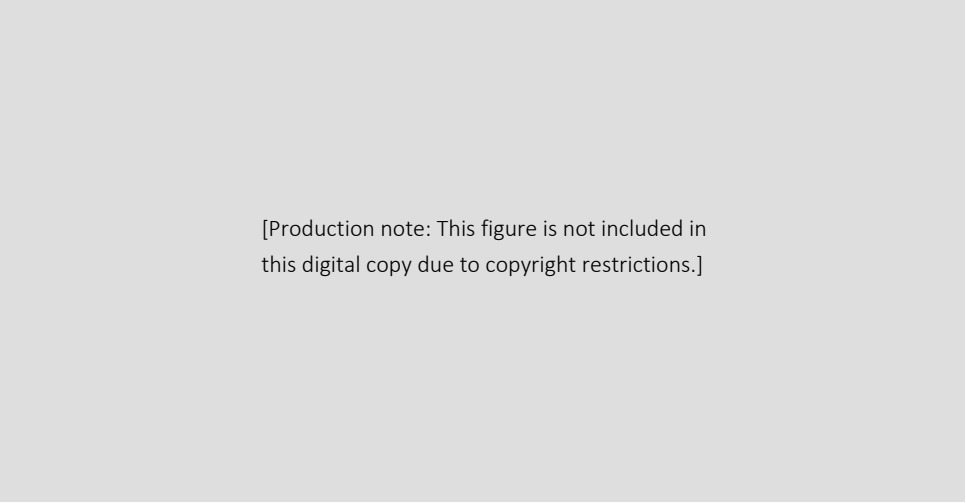


Figure 2.13 : The DEEP_Pachi model block diagram: (a) feature extraction using ensemble backbone models (DenseNet201 and VGG16), (b) patch embedding, (c) position embedding, (d) DEEP_Pachi components—self-attention network and MLP layer, and (e) testing stage with new images on the trained DEEP_Pachi network. *Source: [159].*

Focusing on interpretability and imbalance issues, Shovon *et al.* [144] introducing a DenseNet201-Xception ensemble combined with a threshold-filtered single instance evaluation (SIE) method. This approach achieved high accuracy while improving interpretability, though architectural complexity and dataset noise posed challenges.The architecture lacks adaptive learning components that could further

improve robustness under varied conditions.

Extending the focus on label-free learning, Javed, Mahmood, and Rajpoot [74] proposed Unsupervised Mutual Transformer Learning (UMTL) for label-free WSI classification. UMTL combines a Transformer Pseudo-Label Generator (TPLG) and a Transformer Label Cleaner (TLC) to iteratively refine pseudo-labels, achieving high Area Under the Curve(AUC) scores, including 0.966 on Camelyon16 [20] and 0.975 on TCGA-LC [156]. While classification accuracy measures overall correctness, AUC evaluates how well the model distinguishes between classes, regardless of thresholds. As shown in Fig. 2.14, the model processes WSIs through pseudo-label generation, refinement, and mutual learning. Although achieving high AUC scores, UMTL does not explicitly address contextual relationships or precise localization needs.

This group of studies demonstrates the growing sophistication of hybrid and ensemble learning strategies in breast cancer classification. Early efforts such as Araújo *et al.* [8] and Zhu *et al.* [191] combined CNNs with SVM or dual-branch CNNs to improve classification, though they lacked deeper contextual modeling. To address this, Yan [179] and Srikantamurthy [148], expanded hybridization by fusing CNNs with LSTMs, to capture spatial and sequential dependencies but faced computational challenges. Enhancing spatial focus, Ukwuoma [159] integrated attention mechanisms in an ensemble model, though generalization remained limited. Meanwhile, Shovon [144] and Javed [74] further extended ensemble and hybrid strategies through instance filtering and mutual transformer learning for label-efficient and noise-tolerant classification. Their contributions reflect the potential of hybrid frameworks to tackle real-world issues such as class imbalance, dataset noise, and lack of pixel-level annotations. However, they lack explicit spatial context modeling.

Overall, these studies reflect a gradual evolution in methodological sophistica-



Figure 2.14 : The UMTL framework. The figure illustrates TPLG's pseudo-label generation, TLC's label refinement, and mutual learning for WSI classification. *Source: [74].*

tion—from early CNN-based classifiers to modular attention-enhanced hybrids, recurrent networks, and ensemble frameworks. Earlier models focused on static spatial features, while later works began incorporating modules for spatial focus (attention), sequential dependencies (LSTMs), and interpretability (Grad-CAM). Yet, several limitations persist across the spectrum: most models rely on fixed class structures, lack dynamic context modeling, and struggle with generalizability across datasets. Additionally, hybrid and ensemble architectures often increase model complexity, memory usage, and training time, requiring careful design to avoid overfitting and inefficiencies. Further optimization of hybrid architectures, particularly through scalable attention and cross-resolution fusion, remains necessary to enhance clinical

adaptability.

2.3.3 Feature-based and Context-Aware Models

This category highlights models that enhance classification accuracy by focusing on local structural features and broader spatial context. These approaches integrate nucleus-specific cues, multi-resolution inputs, and semantic-level feature fusion to detect metastasis and distinguish complex subtypes. By improving specificity and reliability, they provide a robust foundation for accurate cancer detection.

This section reviews deep learning methods that focus on advanced feature extraction and early forms of contextual integration for breast cancer classification in histopathology images. While many of these models do not explicitly use recurrent structures, they demonstrate an important shift from isolated, patch-level classification toward architectures that incorporate broader tissue context, spatial structure, or semantic fusion through mechanisms such as multi-resolution inputs, attention, or multi-stage pipelines. These models represent foundational steps toward true context-aware analysis by attempting to capture diagnostic patterns beyond isolated features.

Early feature-based methods focused on local patterns. Zheng *et al.* [189] by proposing a nucleus-guided CNN, achieving 96.4% accuracy on 715 WSIs. This nucleus-centered hierarchical CNN demonstrates effective extraction of local cellular features. However, it lacks global spatial awareness or modeling of inter-regional dependencies. Addressing this limitation, Ehteshami Bejnordi *et al.* [20] introduced a context-aware stacked CNN combining cellular and tissue-level features. This model achieved an AUC of 0.962, but frequently misclassifications between visually similar categories. Their stacked CNN offers one of the earliest forms of spatial context integration. However, both models lacked explicit modeling of spatial relationships across regions.

More advanced designs introduced hierarchical and class-aware learning. Building on early context-aware designs, Han *et al.* [61] introduced the structured CNN model (CSDCNN), improving classification via hierarchical and class-aware feature learning. Their model achieved 93.2% accuracy on the BreaKHis [147] dataset. While improved subtype classification using hierarchical feature learning, but struggled to adapt to unseen subtypes.

The foundational study by Bardou *et al.* [15] explored CNNs for feature extraction, achieving 96.96% binary accuracy on the BreaKHis [147] dataset. However, the model struggled with multi-class classification, and its reliance on large patches increased inference times to 3–5 hours per WSI, demonstrating a trade-off between precision and efficiency.

Multi-modal integration further expanded context modeling. Ektefaie *et al.* [41] advanced feature-based models by integrating multi-omics data, bridging molecular profiles with visual features. Using the TCGA BRCA [79] dataset, the model achieved AUC scores ≥ 0.950 for molecular markers, though small sample sizes and shared phenotype signals limited specificity. This model introduces semantic-level context by associating image morphology with genomic markers. However, it lacked spatial relationship modeling between regions. Hirra *et al.* [65] developed a patch-based model, Pa-DBN-BC, utilizing Deep Belief Networks (DBN) for binary classification. Their approach achieved an accuracy of 84.23% on a dataset from multiple institutions, however, it was limited to binary classification without inter-patch dependency modeling.

Extending feature selection and hierarchical modeling, Murtaza *et al.* [116] developed BMIC_Net, a hierarchical classifier using AlexNet, Information Gain (IG) and Principal Component Analysis (PCA). It achieved over 95% accuracy on BreaKHis [147] dataset,however, the model lacked end-to-end and contextual modeling capabilities.

Deeper CNN models enhanced subtype and grading classification. Zewdie *et al.* [183] incorporated subtype and grading classifications into a ResNet50-based system, addressing a gap in grading. Techniques like CLAHE [192] and histogram matching enhanced the performance, achieving up to 96.75% accuracy across multiple datasets, e.g., BreaKHis [147]. However, its context modeling was limited to specific grading categories.

These studies demonstrate the evolution of feature-based and context-aware models to improve breast cancer classification by combining local structure and broader spatial context. Early works, such as Zheng *et al.* [189] established an early direction in feature-based modeling through detailed local region analysis without capturing the contextual interactions across tissue structures. To reflect broader tissue context, Ehteshami Bejnordi *et al.* [20] improved context representation by combining cellular and tissue-level features, though without explicit inter-region modeling. Han *et al.* [61] advanced hierarchical learning to distinguish subtypes.while Bardou *et al.* [15] focused on binary classification but highlighted challenges in multi-class accuracy and computational costs.Ektefaie *et al.* [41] further extended the context-aware approach by integrating multi-omics data, linking morphological and genetic features.

Patch-based and hierarchical models, such as Murtaza *et al.* [116] added hierarchical structuring through BMIC_Net, combining handcrafted and learned features to improve subtype classification, though its reliance on static feature selection limited adaptability. Zewdie *et al.* [183] further incorporated subtype grading with a ResNet50 model enhanced by preprocessing but lacking spatial modeling capabilities. Finally, Hirra *et al.* [65] improved clinical utility through patch-based and segmented WSI processing but were constrained by binary classification.

Collectively, these models emphasize enhanced feature representation and pre-

liminary spatial awareness, laying the groundwork for more context-aware analysis. However, most architectures remain static and modular, lacking mechanisms to capture inter-patch dependencies or spatial continuity. Several approaches attempt to approximate context through multi-resolution inputs, patch aggregation, or attention mechanisms, yet they fall short in capturing sequential or structural continuity across tissue regions. This highlights an important gap for models capable of context-aware reasoning through spatial flow or inter-patch dynamics.

2.3.4 Adaptable Learning and Model Enhancement Strategies

Adaptable strategies like transfer learning, boosting, and multiple instance learning (MIL) enhance model accuracy and adaptability in breast cancer classification. Boosting addresses class imbalance, MIL detects localized metastatic regions, and transfer learning optimizes performance with limited data. Together, these techniques improve diagnostic sensitivity and accuracy, enabling robust models suited to the complexity of clinical breast cancer diagnostics.

Early efforts focused on transfer learning and boosting methods. Mehra *et al.* [111] evaluated transfer learning for magnification-independent classification, finding that fine-tuned VGG16 achieved 92.60% accuracy and a 95.65% AUC on BreaKHis [147]. While efficient, their approach is sensitive to class imbalance and spatial dependencies. Building on the feature extraction improvements proposed by Mehra *et al.* [111], Vo *et al.* [164] introduced a boosting strategy within a CNN framework to improve class balance and feature extraction from multi-scale images. The model achieved 99.5% binary classification accuracy and up to 96.9% accuracy across magnifications on BreaKHis [147]. However, it lacks mechanisms for effective spatial dependency modeling which is critical for accurately localizing metastatic regions.

In parallel, several works have leveraged MIL to address label scarcity and en-

hance weakly supervised learning. Sudharshan *et al.* [152] addressed label heterogeneity with an MIL framework using patient-level labeling without image-level labels. However, it faced limitations in local pattern representation where precise localization and boundary delineation are crucial. Extending MIL with contrastive learning, Schirris *et al.* [137] advanced MIL with DeepSMILE, integrating contrastive self-supervised learning and variability-aware MIL (Fig. 2.15). Although the DeepSMILE framework effectively employs contrastive learning for feature extraction, it lacks components designed to capture spatial dependencies across WSIs.



Figure 2.15 : Overview of the DeepSMILE workflow: illustrating the contrastive self-supervised pre-training, variability-aware MIL process, and final classification pipeline. *Source: [137].*

More recently, self-learning and hierarchical refinement strategies have been explored to further improve robustness. Abdulaal *et al.* [2] addressed classification challenges with noisy labels by developing a self-learning deep neural network combining transfer learning and hierarchical self-learning. Using ten pre-trained CNNs, including Inception V3 and VGG19, the model iteratively refined label accuracy in

sub-images, achieving 99.1% classification accuracy on BreaKHis [147]. While effective against noisy labels, the approach did not incorporate spatial modeling that can effectively capture complex structural patterns.

Building upon this, Addo *et al.* [4] introduced BCHI-CovNet, a lightweight model leveraging depth-wise separable convolution and second-order pooling (SOP) to handle varying image resolutions efficiently. BCHI-CovNet achieved 99.15% accuracy at 40x magnification and 99.38% on BACH [9]. Despite its efficiency, image color variations affected feature extraction, highlighting the need for future color normalization and higher-order feature integration for improved accuracy. It similarly lacked context-aware mechanisms for complex spatial relationships.

Finally, advancing weakly supervised MIL, Yao *et al.* [180] proposed SoS2MIL, integrating self-supervised learning and second-order pooling to extract complex structural features. Tested on Camelyon16 [20] and a private clinical dataset, SoS2MIL achieved an accuracy of 0.9845 and AUC of 0.9800.The integration of second-order pooling within MIL is innovative, but it remains constrained by the absence of mechanisms to capture sequential or spatial relationships.

Adaptable learning strategies, including transfer learning, boosting, and MIL, have improved classification robustness and efficiency in breast cancer histopathology. Mehra [111] leveraged transfer learning to enhance adaptability, revealing efficiency gains but risks of overfitting due to network sensitivities. Building on this, Vo *et al.* [164] improved class balance using boosting strategies, achieving high accuracy across magnifications, though data scarcity remained a challenge. MIL-based methods advanced weak supervision. Sudharshan *et al.* [152] addressed label efficiency with MIL, improving patient-level classification but lacking detailed image-level insights. Shifting toward self-supervised methods, Schirris *et al.* [137] advanced self-supervised learning with DeepSMILE, reducing label requirements but encountering

variability issues in annotation quality.

Further improvements combined hierarchical learning and lightweight architectures. Abdulaal *et al.* [2] achieved high accuracy under noisy labels using hierarchical self-learning, but computational demands highlighted efficiency needs. To improve deployment efficiency, Addo *et al.* [4] focused on lightweight models with BCHI-CovNet, excelling in resource-limited environments but limited by color variations. Building on this line of research, Yao *et al.* [180] extended weak supervision with SoS2MIL, integrating second-order pooling to capture complex WSI structures, achieving high accuracy but needing broader validation for generalizability.

Table 2.1 provides a comprehensive comparative overview of the reviewed studies on breast cancer classification. It highlights key aspects, including research objectives, methodologies, datasets utilized, reported accuracies, and notable limitations. Although these studies are primarily formulated as classification problems, several implicitly provide coarse localization cues (e.g., attention maps or instance-level scoring), reflecting the inherent connection between slide-level prediction and spatial tumor identification in WSI analysis.

Table 2.1 : Overview of 2016-2024 papers on BC classification in HP images using deep learning.

#	**Authors/Year**	**Objective**	**Approach**	**Dataset**	**Accuracy**	**Key Limitation**	**Category**
1	Bayramoglu *et al.*, 2016 [16]	Develop a system to classify breast cancer images regardless of magnification	Single-task CNN for malignancy; Multi-task CNN for malignancy and magnification level	BreaKHis	83.25% (single-task); 82.13% (multi-task); 80.10% (magnification)	Uneven magnification-specific data; needs better generalization	Multi-Scale and Resolution-Invariant Models
2	Araújo *et al.*, 2017 [8]	Automate classification of breast cancer histology images	CNN for feature extraction + SVM for classification	Bioimaging 2015 breast histology challenge dataset [10]	66.7% (patch-wise, CNN); 77.8% (image-wise, four classes, CNN+SVM); 83.3% (binary, CNN+SVM)	Moderate classification accuracy; SVM adds complexity; Label noise due to region-specific annotation gaps.	Hybrid and Ensemble Learning Architectures
3	Han *et al.*, 2017 [61]	Multi-classify breast cancer subtypes	Structured CNN (CSDCNN) for multi-class classification	BreaKHis	93.2%	High computational cost; large-scale data needed; Predefined classes reduce adaptability; needs dynamic models	Feature-Based and Context-Aware Models
4	Zheng *et al.*, 2017 [189]	Improve feature extraction for lesion classification using nucleus-specific features	Nucleus-guided, three-tier hierarchical CNN	Motic Medical Diagnostic Systems dataset (715 WSIs; 175 annotated)	96.4%	Limited ability to capture non-nucleus-centered information; Lacks tissue-level context; limited for complex structures.	Feature-Based and Context-Aware Models
5	Ehteshami Bejnordi *et al.*, 2017 [20]	Multiclass classification of breast WSIs into normal/benign, DCIS, and IDC	CAS-CNN with cellular and contextual feature learning	221 WSIs of H&E-stained breast tissue from 122 patients	AUC 0.962 (binary); 81.2% (three-class)	Misclassification between DCIS and IDC; high computational cost with larger patch sizes	Feature-Based and Context-Aware Models
6	Bardou, Zhang, & Ahmad, 2018 [15]	Classify breast cancer histology images into benign/malignant and subtypes	CNN with data augmentation vs. handcrafted features + SVM	BreaKHis	96.15% (binary, augmented); 98.33% (binary, ensemble); 83.31% (multiclass, augmented); 88.23% (multiclass, ensemble)	Lower multi-class accuracy; high computational cost	Feature-Based and Context-Aware Models
7	Zhu *et al.*, 2019 [191]	Improve classification by assembling compact CNNs	Hybrid CNN with dual-branch architecture	BreaKHis, BACH	85.2% $\pm$2.6 (patient-level); 87.5% (BACH)	Over-pruning reduces performance; compression not fully optimized	Hybrid and Ensemble Learning Architectures
8	Mehra, 2018 [111]	Compare transfer learning to training from scratch	Fine-tuned VGG16 with logistic regression classifier	BreaKHis	92.6% accuracy; 95.65% AUC	Class bias, network sensitivity, training data dependency	Adaptable Learning and Model Enhancement Strategies

Table 2.1 : Overview of 2016-2024 papers on BC classification in HP images using deep learning. (continued).

#	Authors/Year	Objective	Approach	Dataset	Accuracy	Key Limitation	Category
9	Vo, Nguyen, & Lee, 2019 [164]	Enhance classification accuracy using DCNN and gradient boosting	DCNN + gradient boosting	Bioimaging 2015, BreaKHis	99.5% (Bioimaging, binary); 96.3% (BreaKHis, 100X)	Limited data and class imbalance	Adaptable Learning and Model Enhancement Strategies
10	Sudharshan *et al.*, 2019 [152]	Improve classification by handling variability with MIL	MIL vs. single-instance methods (SIM)	BreaKHis	92.1% (40X)	Requires bag-level labeling; future improvements needed	Adaptable Learning and Model Enhancement Strategies
11	Yan *et al.*, 2020 [179]	Improve classification using hybrid CNN-RNN	Hybrid CNN-RNN model	Dataset of 3,771 histopathology images	91.3% (four-class)	Needs larger datasets, advanced algorithms	Hybrid and Ensemble Learning Architectures
12	Murtaza *et al.*, 2020 [116]	Multi-classify breast cancer subtypes	BMIC_Net with hierarchical classification	BreaKHis	95.48% (binary); 94.62% (subtypes)	Requires large dataset for full optimization	Feature-Based and Context-Aware Models
13	Ektefaie *et al.*, 2021 [41]	Explore relationship between histopathology and multi-omics data	Weakly supervised deep learning	TCGA-BRCA	AUC $\geq$ 0.950	Limited interpretability; sample size constraints	Feature-Based and Context-Aware Models
14	Hirra *et al.*, 2021 [65]	Develop an accurate, patch-based deep learning model (Pa-DBN-BC) for breast cancer detection	Patch-based deep belief network (DBN) for feature extraction	HP images from multiple institutions	84.23%	Limited by hardware resources for extensive input patches; focused on binary classification	Feature-Based and Context-Aware Models
15	Zewdie, Tessema, & Simegn, 2021 [183]	Automate classification of breast cancer types, subtypes, and grade levels	ResNet 50 with data augmentation, resizing, CLAHE	BreakHis, Zendo, Jimma University Medical Center (JUMC)	96.75% (binary); 93.86% (grade)	Limited grade classification due to data constraints	Feature-Based and Context-Aware Models
16	Ukwuoma *et al.*, 2022 [159]	Develop a robust model for multiclass breast cancer classification	Hybrid model combining DenseNet201, VGG16, and self-attention mechanisms	BreakHis, ICIAR 2018	1.0 (benign); 0.99 (malignant)	Further evaluation needed on diverse datasets	Hybrid and Ensemble Learning Architectures
17	Zhou *et al.*, 2022 [190]	Improve efficiency and accuracy in breast cancer classification	ADSVM filtering and RANet for adaptive resolution-based classification	BreaKHis, BACH 2018	98.83% (BreaKHis, patient-level); 99.25% (BACH, binary)	Reduced accuracy at highest magnification	Multi-Scale and Resolution-Invariant Models
18	Schirris *et al.*, 2022 [137]	Improve genomic label classification (MSI and HRD) from WSIs	DeepSMILE with SimCLR-based self-supervised learning + MIL	TCGA-CRC, TCGA-BC	AUROC: 0.87 (MSI); 0.81 (HRD)	Reduced accuracy with noisy data	Adaptable Learning and Model Enhancement Strategies
19	Srikantamurthy *et al.*, 2023 [148]	Classify breast cancer subtypes	Hybrid CNN-LSTM leveraging transfer learning	BreakHis	99% (binary); 92.5% (multi-class)	High computational demands; limited interpretability	Hybrid and Ensemble Learning Architectures
20	Shovon *et al.*, 2023 [144]	Improve HER2 subtype classification	DenseNet201 + Xception with SIE and Grad-CAM	BCI dataset (4870 images)	ACC 97.12% (H&E); 97.56% (IHC)	Class imbalance and noisy data	Hybrid and Ensemble Learning Architectures

Table 2.1 : Overview of 2016-2024 papers on BC classification in HP images using deep learning. (continued).

#	**Authors/Year**	**Objective**	**Approach**	**Dataset**	**Accuracy**	**Key Limitation**	**Category**
21	Taheri and Golrizkhatami, 2023 [153]	Automated diagnosis with magnification-specific and independent models	DenseNet201 with score-level fusion	BreakHis	Best ILA 93.76% (40×); best PLA 96.7% (200×)	Reduced accuracy at 400× magnification	Multi-Scale and Resolution-Invariant Models
22	Javed, Mahmood, & Rajpoot, 2024 [74]	Classify WSIs using unsupervised learning	UMTL with TPLG and TLC for iterative labeling	Camelyon16, TCGA-LC, etc.	AUC 0.966 (Camelyon16); 0.991 (TCGA-RCC)	Sensitive to class imbalance	Hybrid and Ensemble Learning Architectures
23	Xiao *et al.*, 2024 [175]	Classify breast cancer cytopathology images	Inceptionv3 with transfer learning	BreaKHis	¿0.92 accuracy	High computational demands	Multi-Scale and Resolution-Invariant Models
24	Abdulaal *et al.*, 2024 [2]	Improve classification using TL and self-learning	Hierarchical self-learning with 10 pre-trained CNNs	BreakHis	ACC 99.1%	Validation on more datasets needed	Adaptable Learning and Model Enhancement Strategies
25	Addo *et al.*, 2024 [4]	Improve classification with a lightweight model	BCHI-CovNet with depth-wise separable convolution	BreaKHis, BACH	ACC 99.15% (BreaKHis, 40×); 99.38% (BACH)	Requires color normalization	Adaptable Learning and Model Enhancement Strategies
26	Yao *et al.*, 2024 [180]	Combine MIL with second-order learning for WSIs	SoS2MIL for detailed structural info	Camelyon16, clinical dataset	AUC 0.9800; ACC 0.9845	Computational efficiency relies on feature approximation	Adaptable Learning and Model Enhancement Strategies

Overall, while these methods improve accuracy and reduce annotation requirements, most rely on static pipelines with limited contextual modeling. Future work should incorporate sequential and spatial learning techniques to enhance metastasis localization and clinical applicability. Motivated by these limitations, recent research in computational pathology has increasingly explored more expressive architectures capable of capturing long-range dependencies and large-scale contextual representations in whole-slide images. The following section therefore reviews emerging approaches based on foundation models and sequence architectures, which aim to improve representation learning and contextual modeling in WSI analysis.

2.3.5 Foundation Models and Sequence Architectures in Histopathology

Recent advances in computational pathology have introduced *foundation models* that learn broadly transferable representations from large-scale histopathology data and can be adapted to diverse downstream tasks.

UNI demonstrates large-scale self-supervised pretraining for histopathology representation learning and systematically evaluates transferability across multiple computational pathology tasks [32]. CONCH extends this paradigm to a visual–language foundation model by jointly pretraining on histopathology images and biomedical text, enabling cross-modal alignment and improved transfer performance [104]. Virchow presents a large pathology foundation model trained at scale and reports robust generalization across cancer types, including out-of-distribution scenarios [166]. Complementing image-only pretraining, TITAN introduces a multimodal whole-slide foundation model that aligns WSI imagery with pathology reports and synthetic captions for general-purpose slide representation learning [39]. In addition, recent clinical benchmarking studies systematically compare publicly available self-supervised pathology foundation models across multiple diagnostic tasks, highlighting both their potential and current limitations in clinical settings [27].

Beyond foundation models, Vision Transformer (ViT)-based architectures and transformer-style multiple instance learning (MIL) frameworks have become increasingly prevalent for whole-slide image (WSI) analysis due to their ability to model long-range dependencies through token-based interactions. TransMIL introduces a transformer-based MIL formulation to capture correlations among instances within WSI bags and has become a representative baseline for transformer-based WSI classification [141]. In the specific context of lymph node metastasis assessment, recent transformer-based studies investigate micrometastasis detection and CAMELYON16-based evaluation, including MetaTrans, which applies meta-learning to transformer architectures for lymph node micrometastasis detection [70], and ViT-based analyses focusing on explainability and attention behavior in CAMELYON16 classification tasks [130].

More recently, *Mamba-like* selective state-space models (SSMs) have emerged as efficient alternatives to attention mechanisms for long-context sequence modeling. Mamba introduces a selective SSM backbone with linear computational scaling in sequence length, motivating interest in high-resolution and long-sequence applications [56]. While early SSM formulations were not pathology-specific, recent studies have begun integrating Mamba-like architectures into MIL frameworks to better scale WSI analysis to gigapixel resolutions.

MamMIL combines Mamba-style sequence modeling with MIL for WSI representation learning [42]. M3amba further introduces memory-enhanced Mamba mechanisms to mitigate contextual forgetting and improve global dependency modeling in WSI classification [188]. MambaMIL+ extends this line of work by proposing additional strategies to stabilize long-range modeling for gigapixel WSI benchmarks [182]. Collectively, these studies indicate that SSM-based approaches are beginning to be explored in computational pathology; however, evidence for their effectiveness in *precise metastasis localization* remains limited compared with more established

convolutional localization pipelines.

From the perspective of this thesis, many foundation-model and transformer/SSM-based approaches primarily focus on slide-level or bag-level prediction and typically require additional architectural components or post-processing strategies to achieve dense, boundary-aware localization [32, 104, 166, 141, 56].

While these modern architectures demonstrate strong performance in large-scale WSI analysis, they typically rely on substantial computational resources and large training datasets. In contrast, the proposed models in Chapters 3 and 4 are explicitly designed for efficient metastasis detection and localization under weak supervision. Chapter 3 emphasizes an efficient end-to-end architecture that integrates attention mechanisms with regression-guided dense map reconstruction, while Chapter 4 incorporates contextual modeling via Bi-GRUs, multi-scale feature refinement using ASPP, and attention-refined skip connections (CBAM) to improve spatial coherence and boundary delineation.

Table 2.2 summarizes these methodological differences and positions the thesis contributions relative to recent foundation models, ViT-based WSI transformers, and Mamba-like SSM approaches. While these modern architectures are often introduced in the context of WSI-level classification, their design choices directly influence their ability to support metastasis localization and dense prediction. Therefore, discussing localization capability at this stage provides a methodological bridge between classification-focused representation learning and the spatially explicit tasks addressed in the following section.

Table 2.2 : Comparison of modern WSI architectures and the proposed models.

Family / Examples	**Primary goal**	**WSI handling / context modeling**	**Detection / localization capability (typical)**
Foundation models UNI [32], CONCH [104], Virchow [166], TITAN [39]	Transferable representations; broad downstream adaptation	Patch- or slide-level representation learning using large-scale self-supervision; transformer encoders; multimodal alignment in selected models	Primarily slide-level or instance-level prediction; dense, boundary-aware localization typically requires additional task-specific heads or post-processing
ViT / Transformer-based WSI methods TransMIL [141], MetaTrans [70]	WSI-level prediction; metastasis detection variants	Transformer-based aggregation over patch tokens (MIL in TransMIL; WSI-derived patch pipelines in MetaTrans), enabling global token interactions	Localization commonly indirect and coarse (instance attention/heatmaps); pixel-level boundary modeling is not inherent
Mamba-like SSM for WSI MamMIL [42], M3amba [188], MambaMIL+ [182]	Efficient long-context modeling for gigapixel WSI	Selective state-space sequence modeling integrated into WSI pipelines (often MIL-style); favorable scaling with long sequences [56]	Emerging in pathology; evidence mainly for representation/classification so far; dense metastasis localization remains under active exploration
Proposed model (Chapter 3)	Joint detection + localization (weak supervision)	Patch-based dense map reconstruction with spatial and channel attention; regression-guided density estimation; no decoder upsampling	Explicit dense localization map enabling metastasis detection and WSI-level classification within a unified end-to-end framework
Proposed model (Chapter 4)	Joint detection + boundary-refined localization (weak supervision)	Encoder–decoder with Bi-GRU context modeling, ASPP multi-scale refinement, and CBAM-enhanced skip connections	Dense and boundary-aware localization with improved spatial coherence and boundary delineation via attention-refined skip connections

Although these architectures primarily aim to improve representation learning and slide-level prediction, their ability to capture broader contextual relationships also provides a foundation for more spatially explicit analysis of tumor regions. This motivates the subsequent discussion of deep learning approaches designed for metastasis detection, localization, and segmentation.

2.4 Deep Learning Approaches for Metastasis Detection, Localization and Segmentation

Metastasis detection, localization, and segmentation are three interconnected yet distinct tasks in the automated analysis of breast cancer using whole slide histopathology images (WSIs). Metastasis detection refers to the binary classification task of determining whether metastatic tumor cells are present in a WSI. It typically involves assigning a positive or negative label to the entire slide and serves as a preliminary screening step. Localization extends this by pinpointing the approximate regions where metastatic tissue is concentrated. This is often achieved through bounding boxes, region proposals, or attention heatmaps that guide further diagnostic focus, albeit with limited spatial precision. Segmentation, on the other hand, provides a fine-grained, pixel-level delineation of tumor boundaries. It separates metastatic cells from surrounding non-cancerous tissues with high spatial accuracy, enabling detailed morphological assessment. Together, these tasks form a diagnostic pipeline from slide-level prediction to precise tumor mapping, each offering increasing levels of spatial and clinical detail.

In this section, the recent, significant works are categorized into three key areas: 1) Metastasis Detection and Localization, 2) Metastasis Segmentation, and 3) Multi-Task and Hybrid Models for Combined Detection, Localization and Segmentation.

This progression reflects a shift from broad metastasis detection and localization

to precise segmentation and sophisticated multi-task frameworks. By integrating advanced architectures, such as hybrid models and feature-enhancing mechanisms, these approaches address challenges related to efficient processing, contextual awareness, and boundary delineation.

2.4.1 Metastasis Detection and Localization

This subsection focuses on methods aiming at accurately detecting and localizing metastatic regions within WSIs. Techniques in this area often involve patch-based classification, heatmap generation, and multi-scale analysis to achieve robust detection and precise localization of tumor regions. By aggregating patch-level predictions or employing deep learning models with attention mechanisms, these approaches aim to enhance diagnostic accuracy and efficiency. Accurate localization of metastasis is essential for cancer staging, treatment planning, and overall prognosis, highlighting the significance of improving both detection sensitivity and spatial precision.

Early approaches such as Bejnordi *et al.* [19] introduced a multi-scale superpixel classification system combined with spatial clustering for detecting ductal carcinoma in situ (DCIS) in WSIs and achieved an excellent performance. They achieve localization by segmenting epithelial regions and applying spatial clustering to map DCIS lesions within WSIs. However, despite its effectiveness, misclassification in regions with lymphocytic infiltrates and necrotic lesions revealed the model's sensitivity to complex tissue types. Adopting patch-based CNN frameworks and cascaded architectures, Wang *et al.* [169] used GoogLeNet for patch-based classification and heatmap aggregation to detect and localize metastatic cancer in sentinel lymph node biopsies. The model generated heatmaps by aggregating patch-level predictions to provide a coarse localization of metastatic regions. However, the approach required analyzing millions of patches, leading to high computational demand. Additionally, false positives from tumor-like regions highlighted limitations in capturing contex-

tual dependencies. Addressing these computational challenges, Chen *et al.* [30] proposed a cascaded CNN for mitosis detection. As illustrated in Fig. 2.16, the two-stage architecture employs Nc for candidate identification and Nf for refined mitosis detection. By focusing on candidate regions, this cascaded approach reduces processing time while maintaining high sensitivity.



Figure 2.16 : A two-stage architecture of the method proposed in [30], where the first model (Nc) quickly identifies mitosis candidates, and the second model (Nf) refines this by accurately distinguishing mitoses from similar structures, improving both speed and accuracy [30].

To improve robustness and spatial coherence, Liu *et al.* [102] applied CNNs to detect and localize small metastatic tumors in gigapixel pathology images using patch-based classification and heatmap post-processing. Their model achieved an AUC above 97% and a sensitivity of 92.4% on CAMELYON16 [20]. However, errors caused by tissue artifacts and ambiguous regions affected performance consis-

tency. This approach lacked mechanisms to capture contextual dependencies across patches, which are critical for accurate localization.

Employing efficient processing strategies, Lin *et al.* [98] introduced ScanNet, an FCN-based method for lymph node metastasis detection and localization in WSIs. It used a divide-and-conquer strategy and OTSU filtering to reduce processing time while maintaining accuracy. ScanNet achieved an FROC score of 0.8533 and an AUC of 0.9875, outperforming human pathologists.

To further improve efficiency, Lin *et al.* [99] extended their prior work by developing Fast ScanNet, which added anchor layers and asynchronous prefetching. This reduced WSI processing time to under a minute using eight GPUs. However, despite enhanced computational efficiency, the model's focus remained primarily on coarse localization rather than fine-grained segmentation or boundary delineation.

Utilizing handcrafted feature-based CNN architectures, Saha *et al.* [135] developed a supervised model combining a five-layer CNN with handcrafted features for mitosis detection. It achieved an F-score of 90% on datasets like MITOS-ATYPIA-14. However, its reliance on 40× magnification and sensitivity to variations in mitotic shapes limited adaptability to other conditions.

Fauzi *et al.* [45] proposed a CNN for tumor region localization in whole-slide H&E-stained breast carcinoma images. The model was trained on five whole-slide images and tested on ten additional slides, out of a total dataset of 15 WSIs. It produces probabilistic heat maps highlighting tumor regions with Dice coefficients ranging from 0.7 to 0.9. Despite a limited dataset and annotation inconsistencies, the model demonstrates robust localization capabilities, suggesting its applicability in scenarios with constrained data availability.

Building on previous efforts to enhance detection and localization precision, such as the patch-based and heatmap aggregation techniques by Wang *et al.* [169], and

the cascaded architectures by Chen *et al.* [30], Mahmood *et al.* (2020) [108], employs Faster R-CNN and deep CNNs. This approach demonstrates a significant advancement in combining detection and localization within a unified framework. Unlike earlier methods focusing solely on heatmap generation or candidate retrieval, this approach integrates region proposal networks (RPNs) to produce precise bounding boxes. This enhancement improves localization accuracy by providing structured predictions rather than coarse heatmaps. Additionally, its cross-dataset generalization testing on TUPAC16 [163] highlights improved robustness over previous architectures constrained by dataset-specific training.

In another approach, Farajzadeh *et al.* [43] proposed a fully-convolutional residual encoder-decoder network for localizing cancerous nuclei in histopathology images, achieving 98.61% accuracy on an ER+ BCa H&E-stained dataset. The model effectively generalizes to partially annotated datasets through block-wise image segmentation, highlighting its robustness in handling complex tissue structures.

Building on efforts to enhance robustness, Wilm *et al.* [172] tackled domain variability in mitotic figure detection using a domain adversarial training approach. They incorporated a Gradient Reversal Layer (GRL) and a domain classifier to minimize scanner-specific feature clustering, improving generalization across different scanners. Their modified RetinaNet achieved an F-score of 0.7183 on the MIDOG dataset [12], showing strong performance with unseen scanners such as PANORAMIC 1000. However, challenges arose with the Leica GT450 scanner due to illumination and contrast variability, indicating the need for enhanced normalization techniques.

In their recent work, Ortiz *et al.* [120] proposed a patch overlap approach for enhancing cancer detection and localization in histopathological images by applying a 3x3 step 2 average pooling filter to overlapping patches. This method improves

labeling consistency by mitigating isolated misclassifications and adjusting CNN decision thresholds, achieving a 20% correction rate over non-overlapping methods. Tested on the HUP and CINIJ datasets [35], the approach reached an accuracy of 94.6%.

Altogether, early approaches, such as Bejnordi *et al.* [19], employed multi-scale superpixel classification with spatial clustering for DCIS detection, demonstrating high performance but struggling with complex tissue types. Building on feature extraction strategies, Wang *et al.* [169] utilized GoogLeNet with patch-based classification and heatmap aggregation, enhancing coarse localization but at the cost of high computational demand and limited contextual awareness.

To improve efficiency, Chen *et al.* [30] introduced a cascaded CNN framework with a two-stage architecture focused on candidate retrieval followed by refined analysis. This method enhanced processing speed but lacked broader contextual integration. Extending efficiency strategies, Lin *et al.* [98, 99] developed ScanNet and Fast ScanNet, incorporating divide-and-conquer methods and asynchronous prefetching to reduce WSI processing times. However, these models primarily achieved coarse localization, struggling with precise boundary delineation.

Liu *et al.* [102] applied CNNs for detecting small metastatic tumors using patch-based classification and heatmap post-processing. Despite high sensitivity, inconsistency from tissue artifacts and limited contextual modeling reduced reliability. Fauzi *et al.* [45] applied a simpler CNN-based approach for tumor region localization using probabilistic heatmaps, achieving robust results but limited by dataset constraints.

Improving boundary precision and contextual awareness, Saha *et al.* [135] integrated handcrafted features with CNNs for mitosis detection, enhancing precision but remaining restricted by limited scale variability. Extending previous efforts,

Mahmood *et al.* (2020) [108] advanced detection and localization by integrating Faster R-CNN with deep CNNs, achieving improved precision through region proposal networks and demonstrating robustness across datasets.

Farajzadeh *et al.* [43] advanced contextual representation using a fully-convolutional residual encoder-decoder network with block-wise segmentation, achieving high accuracy but facing limitations due to static post-processing and adaptability issues. Wilm *et al.* [172] addressed domain variability through domain adversarial training, enhancing robustness across scanners but highlighting challenges with illumination and contrast variability. Ortiz *et al.* [120] further improved localization accuracy by applying a patch overlap approach with filtering techniques, achieving a 20% correction rate over non-overlapping methods.

These studies collectively illustrate the progression from early superpixel methods to advanced CNN and FCN frameworks, improving diagnostic precision and efficiency. Despite substantial advancements, many methods still struggle with handling complex tissue structures, boundary delineation, and computational scalability. Techniques relying heavily on patch-based processing or handcrafted features face limitations in adaptability and generalization. Additionally, models employing basic CNN architectures often lack mechanisms to capture spatial dependencies across broader tissue contexts. To address these gaps, integrating more advanced feature extraction techniques, multi-scale architectures, and robust domain adaptation strategies can enhance detection and localization performance. Improving robustness to artifacts, optimizing processing efficiency, and expanding validation to diverse datasets are essential steps toward developing clinically reliable diagnostic systems.

2.4.2 Metastasis Segmentation

Metastasis segmentation is essential for accurately delineating cancerous regions within histopathology images, providing detailed structural and morphological information crucial for diagnosis, grading, and treatment planning. Unlike detection or localization alone, segmentation offers pixel-wise precision, which improves the accuracy of downstream tasks like feature extraction and tumor quantification. It is particularly important in WSIs where identifying small metastases and isolated tumor cells (ITCs) is challenging. Additionally, segmentation can enhance localization by providing more precise boundaries, resulting in more accurate and reliable assessments. The following studies discuss various research efforts in metastasis segmentation.

Saha *et al.* [134] developed Her2Net, an attention-based framework for HER2-stained breast cancer images. Combining convolutional and deconvolutional components, Her2Net enhances feature preservation and reduces pixel distortion, ensuring accurate segmentation of cellular structures (Fig. 2.17).

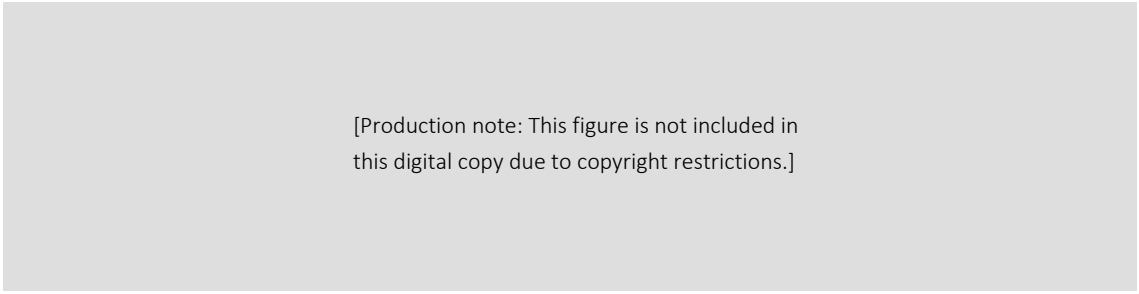


Figure 2.17 : An overview of the two main components of Her2Net: the convolutional section and the deconvolutional section [134].

Her2Net demonstrated the importance of attention mechanisms in histopathology by achieving high segmentation accuracy of 98.33%. However, its dependency on extensively annotated datasets limits its applicability in smaller institutions [134].

To address weakly supervised segmentation, Li *et al.* [93] proposed SegMitos, a model utilizing a concentric loss function to enhance segmentation with minimal labeling. The model expands centroid annotations into concentric circles, effectively capturing mitotic areas in weakly annotated datasets. As illustrated in Fig. 2.18, SegMitos generates pixel-wise segmentation maps and refines them through smoothing and binary operations. A final morphological filtering step reduces false positives, indicated by green circles in the results. The model achieved strong F-scores up to 0.673 across benchmark datasets. The SegMitos model by Li *et al.* [93] effectively enhances segmentation using weakly labeled data by expanding centroid annotations into concentric circles. However, due to the nature of weak supervision, the model's ability to accurately delineate complex structures is inherently limited. While the concentric loss function helps capture broader mitotic areas, it may struggle to differentiate fine-grained details or accurately delineate boundaries, especially in regions with overlapping or irregular tissue structures.

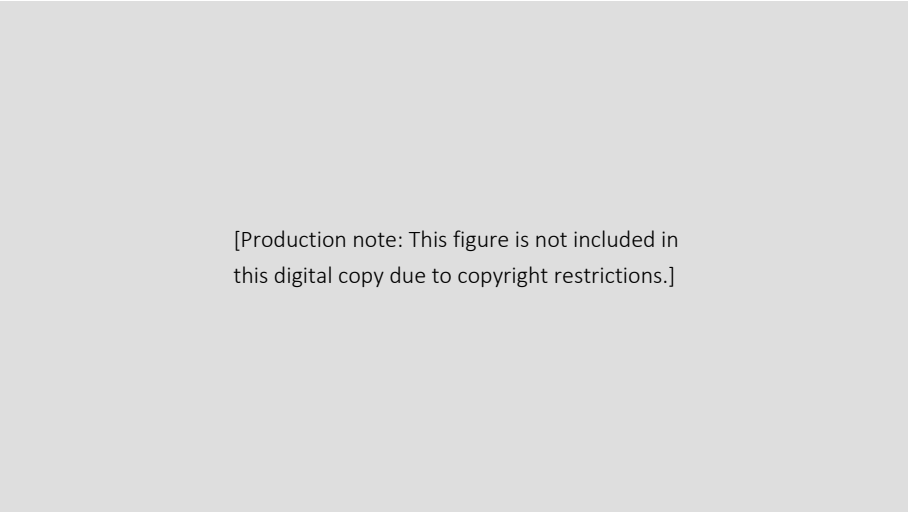


Figure 2.18 : The SegMitos model [93] detects mitosis using concentric labeling, marking the mitotic area and excluding surrounding regions from loss calculation.

Refining tile-wise segmentation approaches, Priego-Torres *et al.* [126] proposed

a DCNN framework for segmenting H&E-stained WSIs. Using tile-wise processing and separable atrous convolutions, it achieved a segmentation accuracy of 95.62% and FWIoU of 92.52% by applying only 12 WSIs. This method reduced tile inconsistencies but was limited by sensitivity to artifacts and reliance on a single scanner. As shown in Fig. 2.19, tile-wise segmentation approaches with (a) direct merging and (b) Dense CRF refinement demonstrate enhanced boundary precision, which is crucial for accurate metastasis segmentation. Although the model achieves strong segmentation performance, its reliance on single-scanner protocols and sensitivity to image artifacts restrict its generalizability.



Figure 2.19 : Tile-wise segmentation approaches explored in the study [126]: (a) direct tile-wise merging of segmentation results, and (b) tile-wise merging using a Conditional Random Field (CRF) for improved accuracy.

Optimizing feature extraction through multi-scale architectures, Schmitz *et al.* [139] introduced msY-Net, a family of U-Net-based FCNs with multi-scale merge blocks

to preserve spatial relationships while reducing memory requirements. The msY-Net models achieved competitive segmentation performance on datasets like PAIP 2019 [85] and CAMELYON 2016 [20] with fewer parameters, making them suitable for resource-limited settings. While msY-Net effectively captures both fine-grained and global features, the reliance on pre-defined scale hierarchies may hinder adaptability.

Incorporating multi-magnification attention mechanisms, Van Rijthoven *et al.* [161] developed HookNet, a dual-branch model that mimics pathologists' multi-magnification diagnostic process. It achieved an F1 score of 0.91 but faced challenges in generalizing across varied fields of view. Despite its ability to capture multi-scale contextual features, HookNet's dependence on specific fields of view can limit its generalizability.

Li *et al.* (2022) [97] proposed the Domain Adaptive Box-supervised Instance Segmentation Network (DBIN) to address challenges in mitosis detection with weakly labeled datasets and domain variability. The model integrates three components: the Box-supervised Instance-Aware (BIA) head, the Pseudo-Mask-supervised Semantic (PMS) head, and the Cross-Domain Adaptive Module (CDAM). While the BIA and PMS heads enhance pixel-wise localization under weak supervision, the CDAM aligns pixel-level features across different domains, improving cross-domain generalization. Evaluated on datasets like MITOSIS2014 [113], MIDOG [12], and TUPAC16 [163], DBIN achieved competitive F1 scores of 0.631, 0.797, and 0.782, respectively.

Focusing on computational efficiency, Khalil *et al.* [81] introduced a modified FCN model for metastatic foci segmentation in WSIs. The model achieved an F1-score of 84.4% and an mIoU of 74.9%. Tested on 188 WSIs, it outperformed eight models, including U-Net and SegNet. It processed each WSI in 2.4 minutes using

four GPUs. While the model offers improved processing speed and segmentation accuracy, its dependence on a single institutional dataset raises concerns about robustness and generalization.

Finally, employing hierarchical segmentation strategies, Xing *et al.* [178] introduced WHRIME, a weight-based recursive hierarchical RIME algorithm [150] designed for multi-threshold segmentation of breast cancer histopathology images. WHRIME addresses the limitations of traditional RIME methods, such as susceptibility to local optima due to low population diversity. It incorporates a hierarchical strategy that divides the population into elite and non-elite groups. This division enhances exploration capabilities and improves convergence accuracy. The algorithm also includes an adaptive weighting factor to balance global and local searches, ensuring optimized segmentation performance. Evaluations on nine images from the Databiox dataset and benchmark functions demonstrated WHRIME's superiority over state-of-the-art methods. The model achieved high metrics, including PSNR and SSIM values of 1.11 and an FSIM of 1.33. These metrics reflect its robust image quality assessment. However, the hierarchical structure introduces computational complexity, which limits scalability in resource-constrained settings. Additionally, WHRIME faces challenges with unimodal and composition functions. Introducing more efficient feature selection techniques or simplifying hierarchical structures could enhance scalability, particularly for real-time analysis of large-scale whole-slide images.

Overall, recent advancements in metastasis segmentation have employed various deep learning architectures, including attention-based frameworks, multi-scale feature fusion, and recursive hierarchical algorithms, to enhance boundary delineation, feature extraction, and efficiency. Techniques like Her2Net [134] leverage attention mechanisms for high segmentation accuracy but require fully annotated datasets, limiting applicability. Weakly supervised models like SegMitos [93] offer efficiency

but struggle with precise boundary delineation.

Building on Tile-wise segmentation methods, [126] achieve high accuracy through separable atrous convolutions and dense CRF refinement but remain sensitive to artifacts and scanner dependency. Expanding upon context-aware modeling, Van Rijthoven *et al.*[161] introduced HookNet, a dual-branch model employing multi-magnification attention mechanisms to capture contextual dependencies, though generalization across varied fields of view remained challenging. Multi-scale architectures such as msY-Net [139] balance fine-grained and global feature extraction but face adaptability challenges due to predefined hierarchies. Building on this groundwork, DBIN [97] presents more sophisticated solutions to cross-domain variability. Furthermore, optimized models like Khalil *et al.*[81] and WHRIME[178] demonstrate efficiency but lack generalization across diverse datasets and face scalability issues.

Despite these advancements, challenges persist with boundary refinement, adaptability to diverse imaging conditions, and robustness across datasets. Models relying on predefined scales or weak supervision often fail to generalize effectively to complex or heterogeneous tissue environments. While these studies demonstrate significant progress in metastasis segmentation through various architectures, including attention mechanisms, multi-scale feature fusion, and optimized CNN frameworks, certain limitations persist. Many models struggle with achieving robust boundary delineation and generalization across diverse imaging conditions. Techniques relying on predefined hierarchies or weak supervision often lack the flexibility to adapt to complex tissue structures or varying staining protocols. Additionally, the reliance on single-institution datasets limits broader applicability. Addressing these gaps requires architectures that can dynamically adapt to heterogeneous imaging conditions, improve boundary refinement, and enhance robustness across diverse clinical scenarios.

2.4.3 Multi-Task and Hybrid Models for Combined Detection, Localization and Segmentation

Multi-task frameworks and hybrid models in breast cancer histopathology have advanced computational efficiency and accuracy. By combining segmentation and classification within a unified framework, these approaches can leverage shared representations and improve generalization across diverse clinical scenarios. They address key challenges like class imbalance, weak supervision, and computational efficiency in large datasets. These advancements streamline workflows and support pathologists in providing faster, more accurate diagnostics.

In an early example of multi-task modeling, Beevi *et al.* [18] introduced a two-stage system combining segmentation and classification within a unified framework. Their Krill Herd (KH) [48]-based localized active contour model (LACM) improved mitotic cell segmentation. A deep belief network (DBN) was then used to classify nuclei as mitotic or non-mitotic. The model achieved F-scores of 84.29% and 75% on the MITOS and RCC datasets. Although effective in improving segmentation accuracy, the model's reliance on contour-based techniques limits its capacity to adapt to irregular boundaries or noisy backgrounds.

Gecer *et al.* [50] laid the groundwork for multi-task deep learning in breast cancer histopathology. Their system classified whole-slide images (WSIs) into five diagnostic categories. It combined an FCN-based saliency detector with a CNN classifier and achieved 55% slide-level accuracy on 240 WSIs, comparable to pathologists. However, computational demands for pixel-wise analysis highlighted scalability challenges. While the approach effectively combines localization and classification tasks, its reliance on pixel-wise saliency detection presents scalability challenges when applied to large histopathology images.

Addressing the challenge of class imbalance in mitosis detection, Wahab *et*

al. [167] proposed TL-HCNN-Mit-Det, a hybrid CNN leveraging transfer learning and data augmentation. By undersampling non-mitotic nuclei and balancing datasets, the model achieved an F-measure of 0.713 on the TUPAC 2016 dataset [163]. Despite its effectiveness, reducing false negatives and improving segmentation accuracy remain open challenges.

Expanding upon domain adaptability, Sebai, Wang, and Al-Fadhli [140] introduced PartMitosis, a partially supervised framework combining weak and strong labels. It integrates two parallel FCNs, one trained on weak centroid labels and the other on fully annotated data. It improves detection accuracy by fusing segmentation maps. As illustrated in Fig. 2.20, a weight transfer function facilitates semantic information sharing between networks, enhancing precision. The model achieved strong F-scores on the ICPR [113] and AMIDA datasets, demonstrating robustness. While PartMitosis effectively combines weak and strong supervision through a dual-network architecture, its reliance on dual-labeling types introduces constraints in achieving robust generalization. The fusion of segmentation maps may struggle to accurately delineate intricate structures or maintain consistency across varying tissue types.

Continuing the advancement of hybrid multitask models, Zhang *et al.* [186] developed DDTNet to detect and segment tumor-infiltrating lymphocytes (TILs). The model used a shared backbone network and multi-scale feature fusion (Fig. 2.21). DDTNet achieved high segmentation accuracy on the TCGA-lym [155] dataset. However, its scalability on large WSIs requires further validation. While the model demonstrates strong generalization capabilities, its application to large WSIs remains limited by the backbone architecture's inability to effectively process high-resolution images.



Figure 2.20 : A detailed overview of the PartMitosis system [140]. The weights of the prediction layer in the strong segmentation branch (W_{Strong}) are derived from the weights in the concentric segmentation branch (W_{Concent}) through a weight transfer function (Γ), which is not learned as a model parameter.



Figure 2.21 : The DDTNet architecture [186], integrating detection and segmentation modules with a feature fusion strategy for precise lymphocyte segmentation.

Furthering the application of multi-task and hybrid models, Khaliliboroujeni *et al.* [82] developed a CAD system for detecting breast cancer metastasis in lymph node WSIs. Using a ResNet50 backbone with spatial and channel attention modules, the model generated density maps to highlight tumor regions. Tested on the CAMELYON16 dataset, it achieved an AUC of 97.03% and an FROC of 84.10%, operating efficiently at 10X magnification to reduce computational load. However, its reliance on limited dataset diversity limits adaptability to complex clinical scenarios. While the ResNet50 backbone and basic attention modules enhance detection efficiency, limitations in capturing complex spatial relationships and contextual details remain. Incorporating advanced attention mechanisms and improved feature refinement through a robust decoder could enhance localization accuracy and boundary delineation, particularly when dealing with subtle or irregular tumor edges.

Building on weak supervision, Li *et al.* [96] proposed SA-MIL, a self-attention-enhanced MIL framework for segmentation without full supervision. As shown in Fig. 2.22, SA-MIL captures global image correlations to overcome the limitations of instance independence in weakly supervised settings. Validated on CAMELYON16, the model achieved an F1 score of 0.868, nearing fully supervised models. This approach demonstrates the potential of combining self-attention and weak supervision to enhance model generalizability, especially when annotated data is scarce. However, higher Hausdorff Distance (HD) values suggest room for improvement in boundary precision.

Incorporating super-resolution with attention mechanisms, Juhong *et al.* [76] proposed a combined SRGAN-ResNeXt and Inception U-net framework. SRGAN-ResNeXt enhances image resolution, enabling the Inception U-net to achieve a Dice similarity coefficient of 0.893 for precise cellular segmentation in H&E-stained images. This workflow, illustrated in Fig. 2.23, demonstrates how super-resolution improves input quality to boost segmentation accuracy. While high-resolution pro-

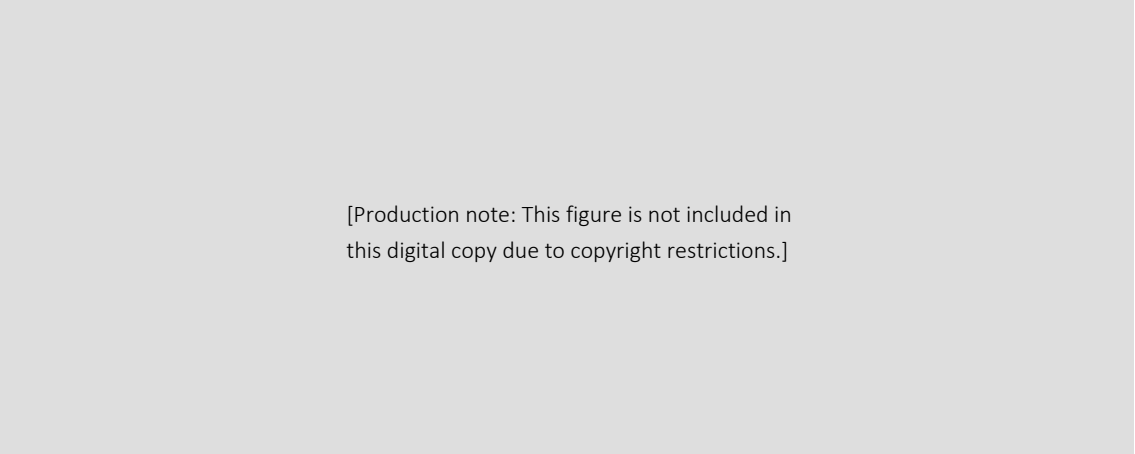


Figure 2.22 : The framework of SA-MIL [96], which uses VGGNet stages and self-attention modules for feature extraction, along with deep supervision and a fusion layer to integrate multi-scale predictions.

cessing enhances segmentation accuracy, the dependency on resource-intensive architectures limits clinical applicability.

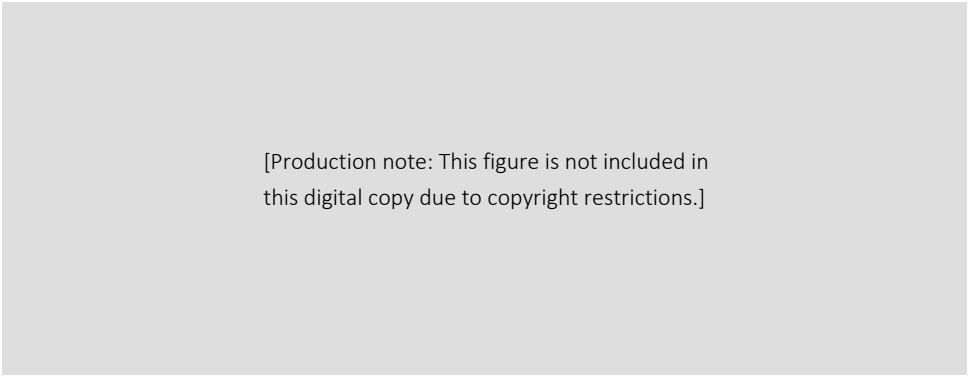


Figure 2.23 : The overall workflow of the models [76]: (a-b) Fresh breast tumor tissues stained and imaged; (c-d) High-resolution images captured using a microscope; (e-g) Low-resolution images simulated and processed by SRGAN-ResNeXt model; (h-i) High-resolution outputs generated; (j-k) Inception U-net Model segments the images.

Leveraging segmentation advances, Jahanifar *et al.* [72] introduced MDFS, a two-stage mitosis detection model inspired by Kahneman's "Thinking, Fast and Slow" theory. The framework combines EUNet for rapid segmentation with EfficientNet-B7 for detailed refinement (Fig. 2.24). To address scanner variability, the model employs domain generalization techniques, including stain normalization. Validated on the MIDOG [12] and ICPR2012 [113] datasets, MDFS achieved high F1 scores. However, challenges remain in detecting faint or small mitoses and minimizing true positive pruning during candidate refinement. Despite its robust performance, the model faces challenges in detecting subtle mitotic figures and managing scanner variability.

More recently, Lu *et al.* [106] introduced BreasTDLUSeg, a two-stage framework for segmenting breast terminal duct lobular units (TDLUs) within WSIs. The first stage uses SPPC-Net to localize TDLU-positive patches, while PPS-Net, a CNN-Transformer model, refines pixel-level segmentation in the identified areas (Fig. 2.25). This coarse-to-fine approach balances accuracy and computational efficiency, achieving Dice coefficients of 79.97.

Despite its success, BreasTDLUSeg's reliance on SPPC-Net's classification accuracy makes it vulnerable to false negatives. Although effective in capturing spatial arrangements and morphological details, the model's reliance on the accuracy of initial detection stages remains a limitation. Enhancing communication between detection and segmentation networks, particularly through refined feature fusion techniques, could improve boundary delineation and robustness in heterogeneous tissue environments.

In closing, the exploration of multi-task and hybrid models reveals notable strides in breast cancer histopathology. Early models like Beevi *et al.* [18] combined active contour segmentation with classification using deep belief networks, demonstrating



Figure 2.24 : An illustration of the mitosis detection method proposed in [72]. (a) data preprocessing, and (b) a two-stage system, 'Detecting Fast' for rapid candidate segmentation and 'Detecting Slow' for detailed refinement of candidates.

reasonable performance but facing difficulties with irregular boundaries and noisy backgrounds.

Expanding on hybrid architectures, Gecer *et al.* [50] combined saliency detection and classification to process WSIs, but scalability issues limited clinical applicability. To address class imbalance, Wahab *et al.* [167] integrated transfer learning and data augmentation, moderately improving mitosis detection precision.

Sebai *et al.* [140] advanced hybrid learning by combining weak and strong labels



Figure 2.25 : The BreasTDLUSeg framework, comprising two main components: (A) SPPC-Net for identifying TDLU-positive patches and (B) PPS-Net for precise pixel-level segmentation within these patches.

within a dual-network structure to improve segmentation accuracy. This approach effectively integrates varied annotations but faces challenges in maintaining robustness across complex tissue structures.

Recent methods have incorporated attention mechanisms to refine segmentation. Zhang *et al.* [186] introduced DDTNet with multi-scale feature fusion for lymphocyte segmentation, while Khaliliboroujeni *et al.* [82] utilized spatial and channel attention modules to enhance tumor detection. Juhong *et al.* [76] further improved accuracy using super-resolution techniques combined with U-net architectures, but computational demands limit clinical applicability.

Li *et al.* [96] addressed weak supervision challenges with SA-MIL, integrating self-attention and multiple instance learning, improving generalizability but facing limitations in boundary precision.

MDFS by Jahanifar *et al.* [72] employed a two-stage model integrating rapid segmentation and detailed refinement, achieving robust performance but facing chal-

lenges with subtle mitoses. Lu *et al.* [106] proposed BreasTDLUSeg, combining localization and segmentation through CNN-Transformer models, achieving high accuracy but struggling with initial classification dependencies.

These models collectively illustrate progress in optimizing diagnostic workflows and addressing key challenges. While recent hybrid models have achieved notable improvements, their ability to accurately delineate complex and irregular boundaries remains limited. Challenges persist in capturing intricate spatial patterns and subtle variations within tumor margins, particularly when dealing with heterogeneous tissue structures. Enhanced feature extraction strategies that effectively capture hierarchical relationships across varying spatial scales could improve segmentation robustness and precision. Developing models capable of refining boundary details and differentiating subtle structures within intricate tissue environments remains essential for achieving higher diagnostic accuracy and broader applicability. Also, future work could explore multimodal integration, adaptive frameworks for diverse imaging conditions, and techniques to minimize annotation dependency while maintaining high diagnostic precision.

Additionally, Table 2.3 provides a comparative summary of these studies. It offers insights into their objectives, methods, datasets, achieved accuracies, and limitations.

Table 2.3 : Overview of 2016-2024 papers on BC detection, localization and segmentation in HP images using deep learning.

#	Authors/Year	Objective	Approach	Dataset	Accuracy	Key Limitation	Category
1	Bejnordi *et al.* (2016) [19]	To detect and localize ductal carcinoma in situ (DCIS) in WSIs using a multi-scale superpixel classification approach.	Multi-scale superpixel classification and region-based classifier with statistical and texture features.	Two datasets with 205 WSIs	Dice score of 0.9243, detecting 80-83% of DCIS lesions with 2-3 false positives per WSI	Misclassification with lymphocytic infiltrates; difficult lesions with large necrotic areas	Metastasis Detection and Localization
2	Chen *et al.* (2016) [30]	Develop a method for mitosis detection and localization using a two-stage CNN approach.	Two-stage CNN model for coarse candidate detection and fine discrimination.	ICPR MITOSIS 2012/2014	F1 score: 0.788 (2012), 0.482 (2014)	Needs larger datasets and GPU optimization	Metastasis Detection and Localization
3	Wang *et al.* (2016) [169]	Identify metastatic breast cancer in sentinel lymph node biopsies.	Patch-based classification and heatmap-based post-processing using GoogLeNet.	CAMELYON16	AUC: 0.925 (whole-slide), 0.7051 (tumor localization)	High computational expense and extensive patch extraction	Metastasis Detection and Localization
4	Beevi *et al.* (2017) [18]	Accurately detect mitotic cells using a multi-classifier system.	KH Algorithm and deep belief network-based multi-classifier system.	MITOS 2014, RCC clinical dataset	F-score: 84.29% (MITOS), 75% (RCC)	Needs more discriminative features and GPU optimization	Multi-Task and Hybrid Models
5	Liu *et al.* (2017)‘[102]	Detect and localize small tumors in gigapixel pathology images.	CNN-based patch classification and heatmap post-processing.	CAMELYON16	AUC: ¿97%	Errors due to tissue artifacts; limited hyperparameter tuning	Metastasis Detection and Localization
6	Fauzi *et al.* (2018) [45]	Tumor region localization in breast carcinoma WSIs.	CNN with patch extraction, model training, tumor probability scoring, and heatmap generation.	15 WSIs (5 for training, 10 for testing)	Dice coefficient: 0.7 - 0.9	Limited dataset size; reliance on basic CNN architectures	Metastasis Detection and Localization
7	Lin *et al.* (2018) [98]	Provide a fast method for detecting lymph node metastasis.	Fully convolutional network (FCN) with a divide-and-conquer strategy.	CAMELYON16	FROC: 0.8533, AUC: 0.9875	Potential improvements with advanced architectures	Metastasis Detection and Localization
8	Gecer *et al.* (2018) [50]	Classify biopsy images into diagnostic categories.	Two-step approach: saliency detection with FCNs, CNN for classification.	240 WSIs	55% slide-level accuracy	Requires refinement for higher accuracy	Multi-Task and Hybrid Models
9	Saha *et al.* (2018) [134]	Automate segmentation of cell membranes and nuclei.	Deep learning framework with convolutional and deconvolutional layers and TLSTM.	HER2 image database	98.33% accuracy	Challenges with cytoplasmic staining and manual annotations	Metastasis Segmentation

Table 2.3 : Overview of 2016-2024 papers on BC detection, localization and segmentation in HP images using deep learning. (continued).

#	Authors/Year	Objective	Approach	Dataset	Accuracy	Key Limitation	Category
10	Saha *et al.* (2018) [135]	Develop an efficient model for mitosis detection and localization.	Supervised deep learning with handcrafted features and CNN.	MITOS-ATYPIA-14, ICPR-2012, AMIDA-13	F-score: 90%	Performance affected by magnification variance	Metastasis Detection and Localization
11	Wahab *et al.* (2019) [167]	Use transfer learning for segmentation and detection of mitoses.	TL-HCNN-Mit-Det model: Pre-trained CNN for segmentation and hybrid CNN for detection.	TUPAC 2016	F-measure: 0.713	Performance affected by segmentation errors	Multi-Task and Hybrid Models
12	Lin *et al.* (2019) [99]	Detect and localizae cancer metastases in large WSIs efficiently.	FCNs and anchor layers for dense scanning.	CAMELYON16	FROC: 0.8533, AUC: 0.9875	Needs large annotated datasets	Metastasis Detection and Localization
13	Li *et al.* (2019) [93]	Develop a weakly supervised method for mitosis detection.	FCN with concentric loss function.	ICPR 2014, AMIDA13, TUPAC16	F-scores: 0.562-0.673	Needs refinement for broader applications	Metastasis Segmentation
14	Mahmood *et al.* (2020) [108]	Automate mitotic cell detection and localization using AI.	Faster R-CNN and deep CNNs (ResNet-50, DenseNet-201)	ICPR 2012, ICPR 2014, TUPAC16	Precision: 0.876, 0.848 (ICPR 2012, 2014)	Needs larger datasets; false positives due to irregular morphology	Metastasis Detection and Localization
15	Sebai, Wang, and Al-Fadhli (2020) [140]	Utilize partially supervised deep learning for mitosis detection.	Two FCNs trained with weak and strong labels; segmentation map fusion with weight transfer.	ICPR 2012, ICPR 2014, AMIDA13	F-scores: 0.575 (2014 ICPR), 0.698 (AMIDA13), 0.788 (ICPR 2012)	Optimization needed for weight transfer; handling class imbalance	Multi-Task and Hybrid Models
16	Priego-Torres *et al.* (2020) [126]	Segment H&E-stained breast cancer images.	Encoder-decoder architecture with separable atrous convolution and dense CRF.	12 WSIs scanned at 40x magnification	Segmentation accuracy: 95.62%, FWIoU: 92.52%	Issues with image artifacts; validation on diverse datasets needed	Metastasis Segmentation
17	Van Rijthoven *et al.* (2021) [161]	Perform multi-resolution semantic segmentation for WSIs.	Two encoder-decoder branches with a hooking mechanism for information fusion.	86 tissue sections	F1 score: 0.91	Performance dependent on fields of view; scaling requires multi-GPU setups	Metastasis Segmentation
18	Schmitz *et al.* (2021) [139]	Segment images at multiple scales, mimicking pathologist workflows.	U-Net-based FCNs with multi-scale merge block.	PAIP 2019, BACH 2020, CAMELYON 2016	Jaccard index: 0.859 (U-Net ensemble)	Needs evaluation on broader datasets	Metastasis Segmentation
19	Farajzadeh *et al.* (2022) [43]	Cancerous nuclei localization in histopathology images.	Fully-convolutional residual encoder-decoder network with post-processing.	ER+ BCa H&E-stained dataset	Accuracy: 98.61%	Dependency on static post-processing; limited generalizability	Metastasis Detection and Localization

Table 2.3 : Overview of 2016-2024 papers on BC detection, localization and segmentation in HP images using deep learning. (continued).

#	Authors/Year	Objective	Approach	Dataset	Accuracy	Key Limitation	Category
20	Zhang *et al.* (2022) [186]	Detect and segment tumor-infiltrating lymphocytes.	Shared backbone network with detection and segmentation modules.	TCGA-lym, BCa-lym, Post-NAT-BRCA	F1 score: 0.739, Dice: 0.717	Limited validation on WSIs; single-expert annotations may introduce bias	Multi-Task and Hybrid Models
21	Li *et al.* (2022) [97]	Improve mitosis detection using weakly labeled datasets and domain adaptation.	Three sub-networks: BIA, PMS, CDAM for cross-domain adaptation.	MITOSIS2014, MIDOG, TUPAC16	F1 scores: 0.631 (MITOSIS2014), 0.797 (MIDOG), 0.782 (TUPAC16)	Time-consuming label expansion; domain adaptation limitations	Metastasis Segmentation
22	Wilm *et al.* (2022) [172]	Address domain variability in mitotic figure detection.	Modified RetinaNet with GRL and domain classifier; ResNet18 backbone.	MIDOG	F-score: 0.7183; 0.7339 (PANORAMIC 1000 scanner)	Scanner-dependent performance; unresolved tissue variation issues	Metastasis Detection and Localization
23	Khalil *et al.* (2022) [81]	Segment metastatic foci in WSIs for diagnosis.	Modified FCN; evaluated on WSIs from National Taiwan University Hospital.	NTUH WSIs	F1: 84.4%, mIoU: 74.9%	Limited generalizability; untested on other cancer types	Metastasis Segmentation
24	Khaliliboroujeni *et al.* (2022) [82]	Detect breast cancer metastasis using WSIs.	ResNet50 with spatial and channel attention; classification and regression.	CAMELYON16	AUC: 97.03, FROC: 84.10	Struggles with hard positive cases at low resolution	Multi-Task and Hybrid Models
25	Juhong *et al.* (2023) [76]	Enhance and segment images in limited technology environments.	SRGAN-ResNeXt for enhancement; Inception U-net for segmentation.	TCGA	IoU: 0.869, Dice: 0.893	Performance depends on dataset quality; high computational demand	Multi-Task and Hybrid Models
26	Li *et al.* (2023) [96]	Enhance segmentation using weakly supervised learning.	Integrates SAMs into the MIL framework for segmentation.	Histopathology datasets (colon, cervix), CAMELYON16	F1: 0.868 (positive), 0.985 (negative)	High Hausdorff Distance; dataset-specific characteristics	Multi-Task and Hybrid Models
27	Jie Xing *et al.* (2024) [178]	Develop a multi-threshold segmentation method addressing local optima issues.	Hierarchical strategy with adaptive weighting for global and local searches.	Databiox (9 images), benchmark functions	PSNR: 1.11, SSIM: 1.11	High computational complexity; challenges in unimodal functions	Metastasis Segmentation
28	Jahanifar *et al.* (2024) [72]	Develop a two-stage framework for efficient mitosis detection inspired by Kahneman's theory.	EUNet for fast detection and EfficientNet-B7 for refined detection.	MIDOG, TUPAC, ICPR2012	F1 score: 0.758 (TUPAC), 0.747 (MIDOG)	Issues with small/faint mitoses; potential false negatives during refinement	Multi-Task and Hybrid Models
29	Lu *et al.* (2024) [106]	Segment TDLUs in histopathological WSIs using a coarse-to-fine approach.	Two-stage SPPC-Net for patch classification and PPS-Net for segmentation.	TCGA-BRCA-TDLU, Post-NAT-BRCA-TDLU	Dice: 79.97%, 92.93%	Performance depends on SPPC-Net accuracy; GPU memory constraints	Multi-Task and Hybrid Models

Table 2.3 : Overview of 2016-2024 papers on BC detection, localization and segmentation in HP images using deep learning. (continued).

#	**Authors/Year**	**Objective**	**Approach**	**Dataset**	**Accuracy**	**Key Limitation**	**Category**
30	Ortiz *et al.* (2024) [**?**]	Enhance detection and localization accuracy using overlapping patches.	Patch overlap approach with average pooling filter, CNN decision thresholding.	HUP, CINIJ (Derived from Cruz-Roa *et al.* [35])	Accuracy: 94.6%	Reliance on handcrafted filtering and static thresholding	Metastasis Detection and Localization

2.5 Summary and Future Work

Deep learning has significantly advanced medical image analysis beyond traditional machine learning. It has enhanced tasks such as classification, detection, localization, segmentation, and feature extraction. By automating image analysis, deep learning has transformed breast cancer diagnosis and improved the effectiveness of Computer-Aided Diagnostic (CAD) systems.

This review began by discussing key medical imaging modalities, including X-ray mammography, MRI, ultrasound, and histopathologic imaging. Each modality's strengths and limitations were analyzed. Histopathology stood out due to its detailed phenotypic insights, which are critical for understanding cancer progression and metastasis.

The review covered influential studies published between 2016 and 2024, offering a comprehensive analysis of deep learning applications in breast cancer histopathology. The reviewed studies were organized into two major research directions: classification models and detection, localization, and segmentation approaches, with a particular emphasis on metastasis identification in whole-slide images (WSIs). In addition, recent advances in foundation models and sequence-based architectures for whole-slide image analysis were discussed, highlighting emerging trends toward large-scale representation learning and improved contextual modeling in computational pathology. While deep learning has enabled remarkable progress, WSIs present ongoing challenges related to high computational demands and the time-intensive nature of manual annotation, which continue to limit large-scale clinical deployment.

Despite the progress, ongoing challenges require further research. A critical issue is the scarcity of fully labeled public datasets of high-resolution histopathology images for effective model training. Creating such datasets is labor-intensive and

expensive, particularly for pathology labs.

Future research could address this limitation by focusing on methods like Few-Shot Learning (FSL), One-Shot Learning (OSL), and Zero-Shot Learning (ZSL). These techniques train models with minimal labeled data and hold promise for breast cancer segmentation and classification, where data acquisition is challenging. Advances in self-supervised learning, weakly supervised learning, and domain adaptation could also enable models to generalize effectively with limited labeled data.

There is also a need for models that deliver accurate results across diverse imaging conditions and modalities. Multimodal fusion, attention-based spatial modeling, and multi-scale feature integration could help integrate data from various imaging sources, improving diagnostic precision and adaptability. Additionally, computational efficiency remains a key challenge. Lightweight architectures and optimization strategies will be essential for real-world clinical applications, especially in resource-constrained environments.

Moreover, several existing models struggle with inter-patch continuity, contextual reasoning, and boundary precision. This highlights an opportunity for novel architectures combining recurrent modules, attention mechanisms, and hierarchical learning to enhance interpretability and localization accuracy.

In conclusion, while deep learning has profoundly impacted breast cancer diagnosis through histopathology images, challenges such as data scarcity, computational efficiency, and cross-modality generalization persist. Addressing these issues can further enhance early diagnosis, personalize treatment planning, and improve patient outcomes in breast cancer care.

Chapter 3

An End-to-End Model for Metastasis Detection in Breast Histopathology Images

Building on the challenges and motivations discussed in previous chapters, this chapter presents a weakly supervised CAD framework for efficient breast cancer metastasis detection and localization in histopathology images. To address the high computational cost of WSI analysis and the limited availability of dense annotations, the proposed framework operates on low-resolution WSIs using minimal supervision. This design reduces computational burden while maintaining competitive detection performance. To maintain computational efficiency, the framework adopts a lightweight architecture without an explicit decoder stage, focusing on direct prediction from encoded feature representations.

Unlike traditional patch-based approaches, the proposed model integrates attention mechanisms with classification and regression tasks to jointly perform WSI classification and tumor localization. A dense localization strategy is employed to generate tumor localization maps and identify individual metastases. Experimental results demonstrate that the proposed approach achieves strong performance under the studied setting, while also providing substantial reductions in processing time and computational cost. These characteristics make the framework suitable for efficient large-scale WSI analysis.

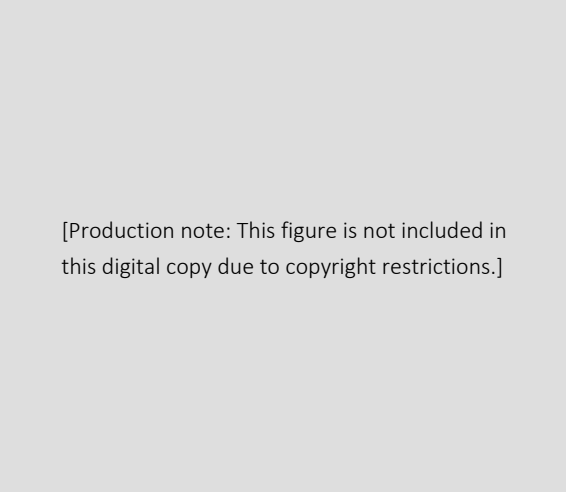


Figure 3.1 : Pyramid representation of a whole slide image, where the level-0 slide (bottom) represents the principal WSI with the highest resolution and the level-n slide presents the smallest image by the base size divided by 2^n [117].

3.1 Introduction

Histopathology plays a critical role in the diagnosis of breast cancer. As mentioned in Section 2.2.4, HP involves analyzing biopsy fragments under a microscope to detect invasive lesions or other anomalies. However, working with histopathology images presents certain complexities. Each slide from these advanced technologies is about one gigabyte. Saving each slide as a pyramidal heap is one solution to having access to several resolutions of one slide. That is why WSIs are usually presented in several levels, from 0 to n. In general, the Level-0 slides represent the principal WSI with the highest resolution and the Level-n slides represent the smallest image by the base size divided by 2^n. In fact, continuing downsampling the slide by dividing its size by two, the same images in several sizes are obtained (Fig. 3.1).

Also, intermittent "noise," as shown in Fig. 3.2, can obscure key details in histopathological images. Additionally, the visual resemblance between healthy and metastatic patches, illustrated in Fig. 3.3, introduces challenges in accurate cancer

detection. The computational complexity and high resource demands of HP imaging further exacerbate these challenges.

To address these challenges, this chapter presents an end-to-end weakly supervised framework for detecting and localizing breast cancer metastases in WSIs.

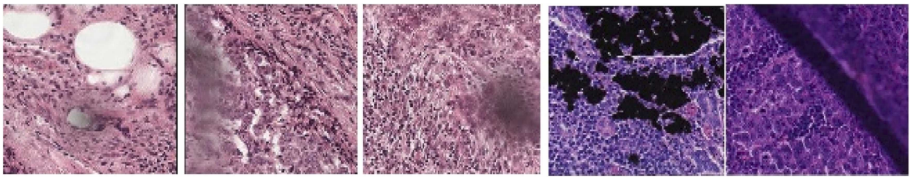


Figure 3.2 : Examples of histopathological image patches exhibiting sporadic noise.

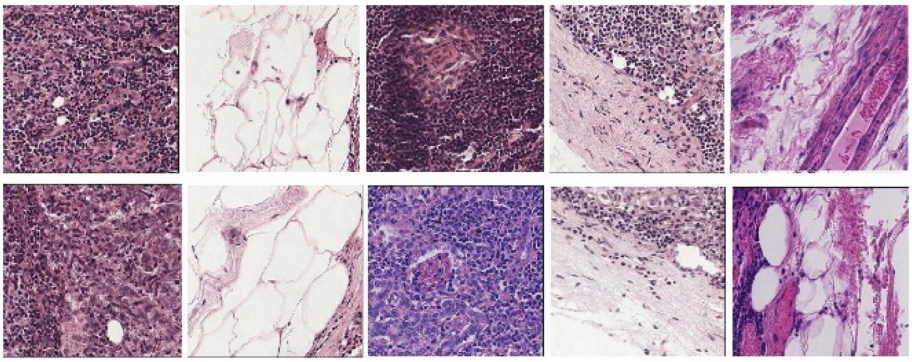


Figure 3.3 : Examples of healthy (the first row) and unhealthy (the bottom row) histopathological image patches. The diversity of tissue structure and sample appearance is observable, highlighting the challenge of distinguishing between normal and tumor patches due to their visual similarities.

The contributions of this chapter are as follows.

- An end-to-end weakly supervised framework is proposed that consists of four main modules: a patch classification module, a tumor ratio regression module, a channel attention module, and a spatial attention module for tumor localization.

- WSI classification and tumor localization are performed simultaneously by extracting dense localization maps while conducting binary classification and tumor ratio regression.

The proposed framework is evaluated on the CAMELYON16 dataset, which was released for the CAMELYON challenge [21]. The method is designed to reduce the computational overhead commonly associated with high-resolution WSI analysis. In addition, computational resources and processing cost are reduced by using an end-to-end architecture that performs classification and localization jointly, without relying on additional procedures such as stain normalization or hard negative and hard positive mining.

Unlike many earlier methods that used the full set of annotated WSIs at 40× resolution, the proposed approach is trained using a reduced subset of the available training data at 10× resolution. Despite this simplified setting, it achieves competitive performance on both WSI classification and lesion localization. The model requires approximately 2.04 minutes on a single GPU to process a WSI for metastasis detection, indicating an efficient computational setting for large-scale WSI analysis.

For the localization task, the proposed model achieved an FROC score [29] of 0.841. This result is slightly higher than the score reported for the CAMELYON16 challenge winner [169], which reported an FROC score of 0.807 using the full set of 270 annotated WSIs at 40× resolution for training. Under the studied weakly supervised setting, these results highlight the potential of the proposed framework for efficient metastasis localization using reduced data and lower magnification.

The rest of this chapter is organized as follows. Section 3.2 reviews the most relevant related work. Section 3.3 presents the proposed approach. Section 3.4 reports the experimental results. Finally, Section 3.5 concludes the chapter.

3.2 Comparison to Preceding Metastasis Detection Approaches

This section briefly contrasts the proposed framework with representative approaches for metastasis detection in whole-slide images (WSIs), focusing on computational design, supervision requirements, and spatial output characteristics.

Early high-performing methods on the CAMELYON16 benchmark [21] primarily rely on dense patch-level classification followed by heatmap reconstruction. For instance, Wang *et al.* [169] employed an ensemble of GoogLeNet models with stain normalization and hard negative mining, while Liu *et al.* [102] used CNN-based patch classification to generate tumor probability maps. Although these approaches achieve strong detection performance, they often involve computationally intensive high-resolution patch extraction and subsequent heatmap reconstruction steps, which limit scalability for large WSIs.

To incorporate spatial context, Kong *et al.* [87] introduced recurrent neural network-based modeling (e.g., LSTM) to capture dependencies between neighboring patches, demonstrating that contextual interactions can improve detection performance. However, such methods typically operate on pre-extracted patch sequences and remain computationally demanding.

Other studies have focused on improving spatial precision through segmentation-based frameworks. For example, Guo *et al.* [57] proposed a two-stage approach combining patch classification with semantic segmentation to refine tumor regions. While this design aims to improve delineation, it requires pixel-level annotations and introduces additional supervision and pipeline complexity.

In contrast, the proposed framework is designed as an efficient end-to-end approach that operates directly on lower-resolution inputs ($10\times$) under weak supervision. Rather than relying on high-resolution patch scanning or multi-stage pipelines, the model jointly performs metastasis detection and localization within a unified

architecture. This design prioritizes computational efficiency and reduced annotation burden while maintaining competitive localization performance, making it more suitable for scalable WSI analysis.

3.3 The Proposed Method

Two main objectives are addressed in this chapter. The first is to estimate the location of metastases within a WSI, and the second is to determine whether metastasis is present in the WSI. In this work, a computationally efficient end-to-end model for WSI classification and metastasis localization is developed. By combining channel attention, spatial attention, patch classification, and a tumor ratio regression module, the proposed framework simultaneously performs whole-slide classification and tumor localization.

Tumor regions are highlighted by reconstructing a dense localization map for each WSI. Unlike typical fully convolutional architectures that rely on a decoding pathway with upsampling layers, the proposed model omits an explicit decoder. This design choice reduces computational complexity and enables faster inference compared with conventional patch-based pipelines, while still producing meaningful localization cues.

Most previous works, *e.g.*, [169, 21, 102], at this stage design separate models to classify WSIs. At first, they design a model for the metastasis localization task and reconstruct tumor heatmaps for the input WSIs.Then, to recognize whether a WSI contains metastasis or not, they develop another model and train this second model using features extracted from the predicted heatmaps of the first model. In the proposed approach, instead, it is not necessary to develop an additional classification model for WSI classification. The metastatic WSIs are identified by aggregating the responses of the detected regions in the reconstructed dense localization map.

The proposed model is built upon a ResNet50 backbone network. The average pooling layer and the final fully connected layer of ResNet50 are removed, and the proposed architecture is attached to the last convolutional layer of the backbone network. In Fig. 3.4, an overview of the model is illustrated. As shown in Fig. 3.4, after the last convolutional layer of ResNet50, the model is divided into two sub-branches. The first sub-branch starts with a spatial module to extract metastatic sections of the input image. The second sub-branch begins with the channel attention module to learn what is noteworthy in the input image. Also, these two modules are connected to boost the information of the extracted features. Subsequently, a classification module for identifying healthy inputs from unhealthy, and a regression module to measure the proportion of the metastatic parts in an input image, are linked to both sub-branches.

In this section, at first, the details of the network architecture of the proposed model are presented. Then, details of the loss functions used in this model are provided.

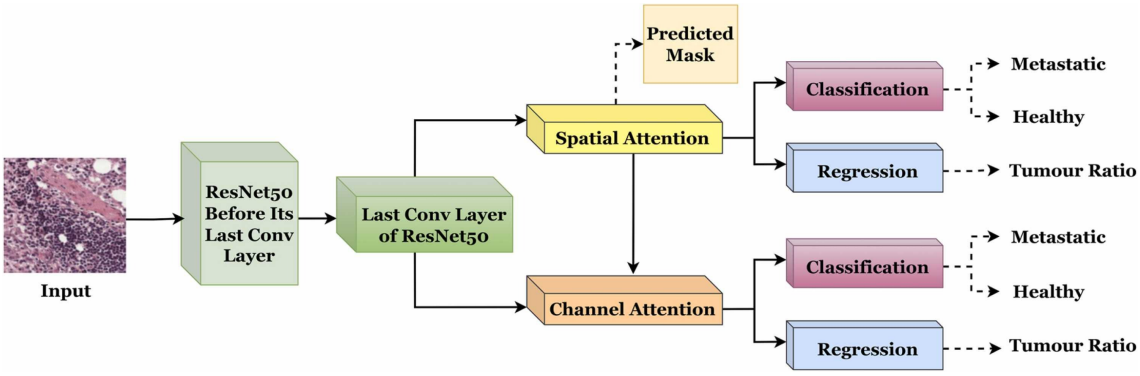


Figure 3.4 : An overview of the model.

3.3.1 Network Architecture

As shown in Fig. 3.4, this model is built based on a ResNet50 backbone network. The proposed structure attached to the backbone ResNet50 is composed of six mod-

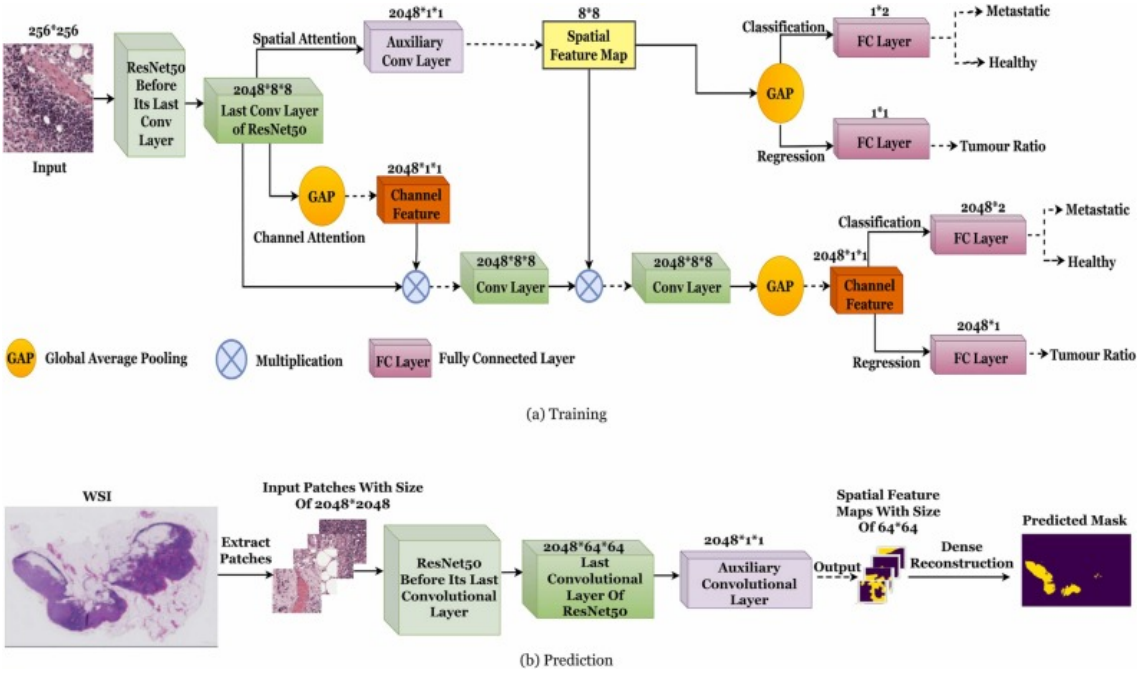


Figure 3.5 : The framework of the model for training (a) and prediction (b).

ules: one spatial attention module, one channel attention module, two classification modules, and two regression modules. The spatial attention module is the Localization branch that produces a density map that highlights the tumor part of each input. The channel attention module extracts the inner features of the channels of the convolutional layer and assists the model in learning meaningful sections of the input image. The Classification sub-branches determine whether an input image is healthy or not, and the Regression sub-branches infer the tumor ratio in the input image. Ultimately, all the modules in this approach assist each other in detecting metastasis more efficiently. By designing these sub-branches, without needing pixel annotation, this model can simultaneously learn whether an input patch is healthy or not, the ratio of the metastases in the patch, and the approximate spatial location of metastatic regions. The model is trained by considering that the elicited dense map should have the same class and tumor ratio as the input image.

In Fig. 3.5, an illustration is presented showing the proposed model's structure

in more detail. Note that by going more deeply through the layers of the CNNs, more semantically meaningful features are learned from the input images. After the last convolutional layer in ResNet50, the model is split into two branches where one branch is started with the spatial attention module and another with the channel attention module. In the spatial attention module, an auxiliary convolutional layer of the size of $2048 \times 1 \times 1$, regardless of the input dimensions, is attached to the last CNN layer. This auxiliary layer extracts the most informative spatial features of the input image by heading the information from the 2048 different channels of the previous convolutional layer into one channel. This branch forms the localization pathway, producing a spatial feature map that highlights potential tumor regions within the input image.

The default input size of ResNet50 is $224 \times 224 \times 3$, and the size of the last convolutional layer at ResNet50 is $7 \times 7 \times 2048$, which varies with the change of the input size. In this model, the input size for training is 256×256, and 2048×2048 in the testing phase. So, the size of the last convolutional layer at ResNet50 in the training phase becomes $8 \times 8 \times 2048$ and $64 \times 64 \times 2048$ in the test phase. Because of the input size, the feature map size obtained in training is 8×8 and in testing is 64×64.

In parallel, the channel attention module focuses on the inter-channel connection of features. By linking a global average pooling layer to the last convolutional layer of ResNet50, a $2048 \times 1 \times 1$ channel feature is produced. Then, by multiplying the produced channel feature and the obtained spatial feature in the last convolutional layer of the backbone ResNet50, a new convolutional layer containing more meaningful features is created. After that, by applying another global average pooling, a more valuable channel feature is produced. Finally, binary classification and tumor ratio prediction are performed by adding average pooling followed by full connection layers for both branches. Instead of considering a pixel-wise loss that needs pixel

annotation, a combination of losses of classification and tumor ratio prediction is considered. The proportion of tumor cells relative to the whole lesion region cells is defined as the tumor ratio. Since the constructed dense map of each input should have the same tumor ratio and class label as the input image, the model is trained for classification and regression. This weakly supervised formulation encourages the dense localization map to be consistent with the patch-level class label and the estimated tumor proportion of the input patch. As a result, the model can produce spatial activation patterns without requiring pixel-level annotations.

3.3.2 Loss Functions

The existing metastasis detection and localization methods in WSIs have generally been presented in two main categories. Patch-wise classification methods train models by learning each patch's label and then producing a heatmap of the final result. They typically apply a patch-level classification loss function such as cross-entropy loss to train the model [169, 21, 102]. Besides, other studies, including methods that apply FCN, achieve the goals based on pixel-wise learning, *e.g.*, [57]. These approaches rely on pixel-wise loss functions that require dense pixel annotations, which are costly and time-consuming to obtain.

In this model, a loss function is considered separately for each branch, including classification and regression branches. So, all branches learn individually and collectively to optimize the model in the right direction. These losses guide the optimization of each sub-branch while jointly contributing to the overall model training. For simplicity, equal weights are assigned to each loss component during training.

As is seen in Fig. 3.5(a), an auxiliary convolutional layer is linked to the last layer of the backbone ResNet50 model, to extract the spatial feature map for each patch. Unlike other approaches, pixel loss is not used in this stage. Instead, two modules corresponding to classification and tumor ratio regression are optimized.

The cross-entropy loss L_{clf1} and the Minimum Square Error (MSE) loss L_{reg1} are considered as the loss functions for classification and regression, respectively.

In parallel, the average pooling layer following the last CNN layer of ResNet50 extracts channel-level feature representations in the second branch, which is also followed by classification and regression operations. Similarly, the cross-entropy loss L_{clf2} and the MSE loss L_{reg2} are considered as the loss functions for classification and regression of this channel-attention branch, respectively.

Eq. 3.1 gives the details of how the cross-entropy loss L_{clf} is calculated for both branches:

$$L_{\text{clf}} = -\frac{1}{N}\sum_{i=1}^{N}\left(t_i \cdot \log(p_i) + (1-t_i)\cdot \log(1-p_i)\right), \tag{3.1}$$

where t denotes the target label, p represents the predicted probability, and N is the batch size.

MSE loss is calculated as shown in Eq. 3.2 as:

$$L_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2, \tag{3.2}$$

where Y represents the ground-truth tumor ratio, $\hat{Y}$ denotes the predicted tumor ratio and N is the batch size.

The model is trained using four loss components where the total loss of the model L_{T} is calculated by summing up the four individual losses as:

$$L_{\text{T}} = L_{\text{clf1}} + L_{\text{reg1}} + L_{\text{clf2}} + L_{\text{reg2}}. \tag{3.3}$$

In Section 3.4.4, the ablation studies have been carried out. The contribution of the spatial and channel sub-branches is examined through the ablation study results. Therefore, as shown in Table 3.2, the experiments show that incorporating both spatial and channel branches provides better performance by considering these four losses in both classification and regression sub-branches compared to only using

one sub-branch (removing Clf_1 and Reg_1) and applying only two losses. It is also shown in Table 3.3 the performance achieved by the model, in comparison to other methods that have only applied patch-level classification, *e.g.*, [169, 21, 102].

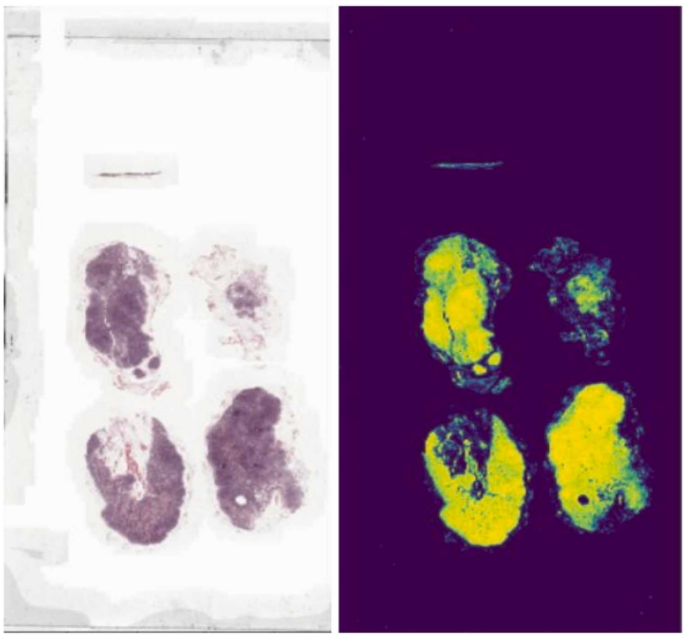


Figure 3.6 : An example of applying the Otsu method on a WSI. The non-tissue background is filtered using the Otsu thresholding method [121].

3.3.3 Training and Inference

The goal of this work is to detect breast cancer metastases in histopathology images that typically contain gigapixel-scale resolution. As WSIs are massive images, before feeding them to the model, they are first divided into smaller patches. Each metastatic WSI comprises both healthy and unhealthy areas. For training the model, a subset of 50 metastatic WSIs and 30 normal WSIs is randomly selected from the CAMELYON16 dataset. To improve computational efficiency, non-tissue background regions are removed using the Otsu thresholding algorithm [121]. The Otsu method automatically estimates a threshold that separates foreground tissue from background regions. All pixels with intensity higher than Otsu's threshold are

considered as the foreground and the rest are remarked as the background. Eventually, a binary mask will be produced where the foreground's pixels take the value of one and other pixels zero. This step preserves the informative tissue regions while removing large background areas. In Fig. 3.6, an example of applying the Otsu method on a WSI is shown.

In the experiments, 710k patches of a size of 256×256 are randomly sampled from these 80 images, where 355k of them include metastasis and 355k do not. Tumor ratio targets are computed offline from the available annotation masks by measuring the proportion of tumor pixels within each extracted patch. These ratios are used only as regression targets during training, and the model itself is trained without pixel-wise supervision.

For each input patch, five different outputs are produced in each repetition, including an 8×8 spatial feature map from the added auxiliary convolutional layer, two classification results looking for the same result, and two regression results that also look for the same outcome.

During inference, image regions of size 2048×2048 are used as model inputs. This is much bigger than the patches used for training and is only restricted by the GPU memory. After training the model, for each WSI, all ROIs of a size of 2048×2048 are sequentially extracted and fed to the model with an overlap of 128 pixels. For each ROI, the model produces a 64×64 spatial probability map (i.e., the output resolution is 1/32 of the input in each spatial dimension). The ROI-level probability maps are merged in sequence to form a whole-slide probability map. Each pixel in the reconstructed map represents the predicted probability of metastasis at the model output resolution.

To obtain a binary mask for region localization, a fixed probability threshold of 0.5 is applied to the whole-slide probability map at $10\times$. This converts the

probability map into a binary mask suitable for deterministic post-processing.

Morphological opening and closing operations are then applied to regularize the predicted WSI masks while preserving the main structural characteristics of detected regions. In this thesis, opening is implemented as erosion followed by dilation, and it is used to suppress small isolated detections and smooth object boundaries. Closing is implemented as dilation followed by erosion, and it is used to fill small holes and connect fragmented regions.

Throughout all experiments in this chapter, opening is applied using a 3×3 square structuring element, followed by closing using a 5×5 square structuring element. These values were selected empirically based on validation data to balance noise removal and preservation of tumor regions, and were then kept fixed across all experiments in Chapter 3. All post-processing steps are applied deterministically to the model outputs, and no post-processing parameters were tuned or adjusted based on test-set performance.

In Fig. 3.7, an example of the predicted mask before and after applying the morphological operations is shown. The zoomed-in views of the cross-sectional areas show that some tiny detected points are removed after applying morphology operations. Therefore, while maintaining the main features in the masks, these operations remove small artifacts while preserving the main detected regions.

Finally, for evaluation, connected components are defined from the post-processed binary mask obtained after thresholding and morphological refinement. For lesion-level localization, each detected region is assigned a confidence score equal to the maximum probability value within that region, taken from the corresponding raw probability map. This max-based aggregation follows standard CAMELYON-style practice for lesion ranking. While it may be sensitive to isolated high-probability pixels, its robustness is further examined in Chapter 4 through a sensitivity analy-

sis using alternative aggregation strategies, which demonstrate comparable performance trends. For WSI-level classification, each predicted WSI probability map is scored by averaging the probabilities over the detected foreground regions, providing a single metastasis score for slide-level evaluation. This separation ensures that post-processing affects only the spatial extent of detections, while confidence values remain determined by the underlying probability map.

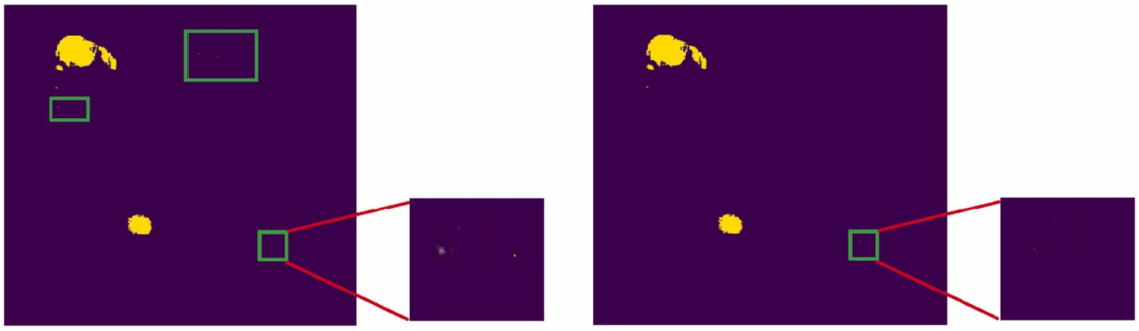


Figure 3.7 : An example of a predicted mask before (left) and after (right) applying the morphological operations on the predicted mask. The green boxes surround some tiny detected points that are hardly visible. The zoomed-in views of the cross-sectional areas show that some small detected spots are removed after applying the morphology operations.

3.4 Experiments

To validate the performance of the proposed approach, experiments are conducted on the CAMELYON16 dataset, a benchmark dataset widely used for breast cancer metastasis detection in histopathology images. In this section, the dataset is first introduced. Then, the evaluation metrics and experiment results are presented.

3.4.1 Dataset

The CAMELYON16 challenge [21] was the first challenge to evaluate metastasis detection using WSIs instead of predefined ROIs. The objective of the challenge was

to detect metastases in H&E stained WSIs of breast cancer lymph node sections. For running this challenge, a large annotated dataset of 400 WSIs from two institutions, *i.e.*, Radboud UMC, and UMC Utrecht, was collected in total. Among the 400 WSIs, 270 WSIs were for training and 130 WSIs were for testing. Additional details of the CAMELYON16 dataset are summarized in Table 3.1.

In this chapter, a subset of 80 WSIs (50 metastatic and 30 healthy) from the CAMELYON16 training set is used to generate training patches. This subset corresponds to less than one-third of the 270 WSIs provided in the official CAMELYON16 training set. The trained model is evaluated on the entire set of 130 CAMELYON16 test WSIs, following the standard evaluation protocol used in previous studies. The data collected from the two institutions are mixed together. This setting allows the evaluation of the proposed framework when trained using a limited subset of annotated training data.

Training patches are extracted from these 80 WSIs. The extracted patches are randomly divided into training (75%) and validation (25%) sets for model development.

Table 3.1 : Details of datasets, the numbers of images used for training and testing in the CAMELYON16 dataset (Bejnordi textitet al.) [21].

Institution	Train (Normal)	Train (Tumor)	Test
Radboud UMC	90	70	80
UMC Utrecht	70	40	50

3.4.2 Evaluation Metrics

The CAMELYON16 challenge [21] defines two evaluation tasks for metastasis detection in lymph nodes among women with breast cancer. The first is tumor

region localization, and the second is WSI-level classification.The FROC curve [29] has been considered to evaluate the performance of the localization task. The FROC curve plots sensitivity versus the average number of false positives in each WSI across the different thresholds. To compare with other works, the average sensitivity at six different false positive points, *i.e.*, 1/4, 1/2, 1, 2, 4, 8 in the FROC, is considered as the final score of the localization task. The area under the Receiver Operating Characteristic (ROC) curve [193], *i.e.*, AUC score, is considered for evaluating WSI classification.

3.4.3 Qualitative Evaluation

At first, the visualization of the experiment results is provided. Examples of metastasis detection results produced by the proposed method in comparison with the ground truth of each whole slide image are shown in Fig. 3.8. The first column shows the whole slide images with the detected tumor polygons drawn on them; the second column shows the ground truth, that is the tumor region in WSI; and the third column shows the predictions generated by the proposed model.

Also, in Fig. 3.8, to illustrate how the model distinguishes in differentiating tumor regions from confusing normal regions, some visually challenging regions of metastatic and healthy areas are enclosed with white and red boxes to highlight them.

3.4.4 Ablation Studies

The ablation studies are conducted to examine the contribution of the channel attention module and the classification and regression modules in the first and second sub-branches. It is noteworthy that the spatial attention module provides the main output (predicted mask) of the model, as shown in Fig. 3.4 and Fig. 3.5(b). The aim of this section is to analyze how these modules influence the model outputs. In

the following, more details of the two ablation studies are provided.

3.4.4.1 The Effectiveness of the Spatial Attention Sub-Branch

In this model, as shown in Fig. 3.5(a), in the first sub-branch, the spatial attention module extracts the spatial feature map. In this stage, as mentioned in Section 3.3.1, rather than a pixel-wise loss, which is an expensive and tedious process for pathologists as it needs pixel annotation, a composition of classification and regression loss functions is applied. Therefore, to evaluate the spatial feature map as output, as placed in Section 3.3.2, the classification module by the cross-entropy loss (L_{clf1}) and the regression module by the Minimum Square Error (MSE) loss (L_{reg1}) are assessed. The model is encouraged to learn spatial activation patterns by learning the class (*i.e.*, cross-entropy loss) and the tumor ratio (*i.e.*, MSE loss).

In this section, an ablation study is conducted to demonstrate the usefulness of the existence of classification and regression modules at the first sub-branch in the model. For this experiment, the model is trained and evaluated without the classification and regression modules of the Spatial Attention Sub-branch. The dataset and all other training parameters remain identical to those described in Section 3.3.3. The results are provided in Table 3.2, where Clf_1, Reg_1, and Spt_Atn, respectively, represent the classification, regression and spatial attention modules in the first sub-branch, and Clf_2, Reg_2, and Chl_Atn consecutively represent the classification, regression and channel attention modules in the second sub-branch. The results obtained without Clf_1 and Reg_1 are lower than those obtained with these modules enabled. It is evident from the table that the FROC score for the region localization task decreases from 0.8410 to 0.8331, and also for WSI classification decreases from 0.9703 to 0.9601.

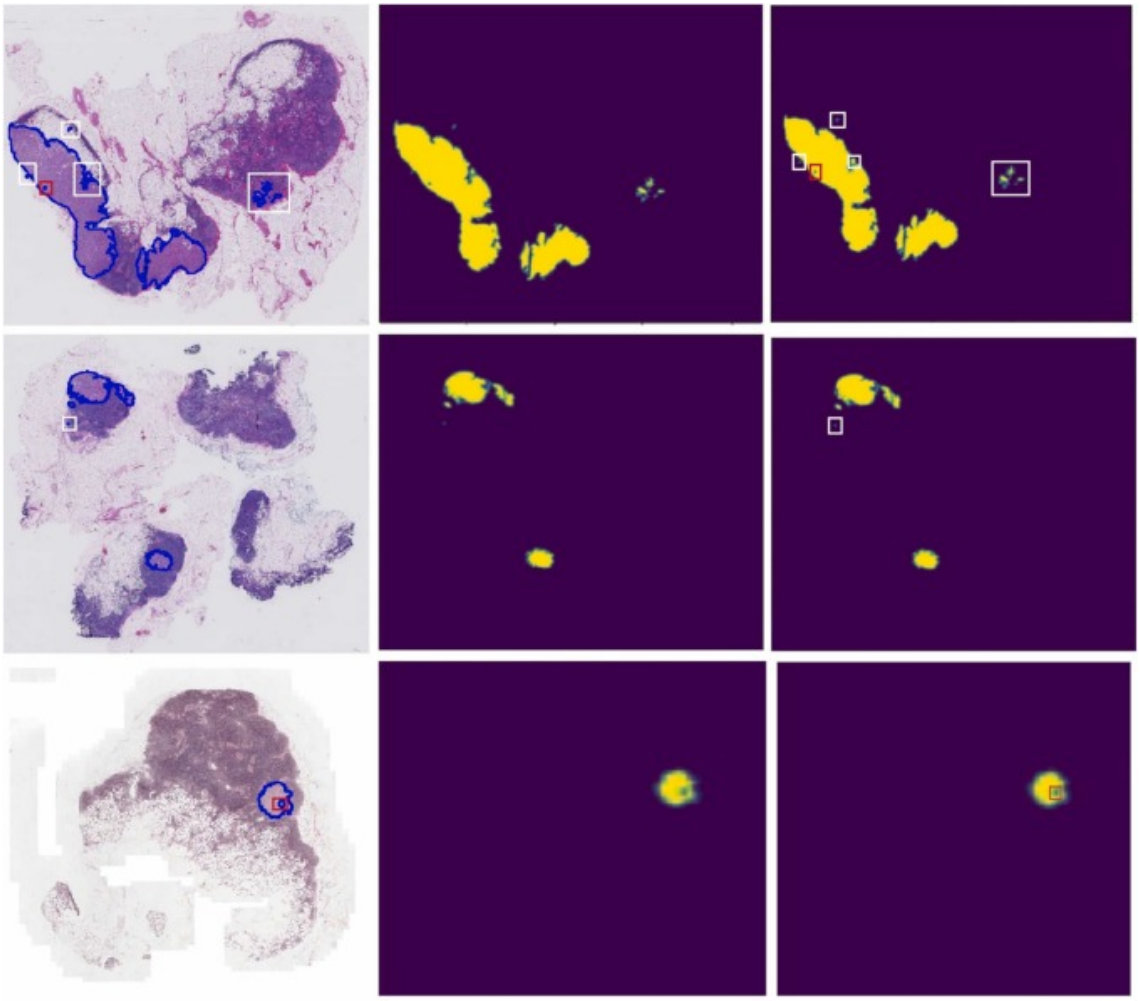


Figure 3.8 : (a) The left column shows the whole slide images where tumor polygons are drawn over them (these contours are indicated in blue), (b) the middle column shows the ground truth of tumor regions in the WSIs, and (c) the third column shows the output of the method. The white boxes show some small tumor areas that are not easy to detect. This method still detects them explicitly. The red boxes show some tricky, healthy spots that also are identified accurately.

3.4.4.2 The Effectiveness of the Channel Attention Sub-Branch

In this section, another ablation study is conducted to evaluate the contribution of the model's channel attention module, followed by classification and regression modules at the second sub-branch. Due to that, the channel attention, classification,

Table 3.2 : Comparison of the results with and without specific modules.

Spt_Atn	Chl_Atn	Clf_1	Reg_1	Clf_2	Reg_2	WSI classification (AUC score)	Region localization (FROC score)
✓	✓	✗	✗	✓	✓	96.01	83.31
✓	✗	✓	✓	✗	✗	95.72	83.02
✓	✓	✓	✓	✓	✓	97.03	84.10

and regression modules at the second sub-branch are removed from the model, and the model is trained and tested while only consisting of the first sub-branch. The same as the ablation study in Section 3.4.4.1, the dataset and all other needed parameters are as specified before in the paper. Table 3.2 indicates that the results obtained without Chl_Atn, Clf_2, and Reg_2 are lower than those obtained with these modules included. It is demonstrated that the FROC score for the region localization task falls from 0.8410 to 0.8302 and for WSI classification falls from 0.9703 to 0.9572.

As shown in Fig. 3.5(b), in the predicting phase, only the spatial attention branch is required. Therefore, the estimated time of processing a WSI in the predicting phase is not dependent on other modules and remains unchanged when removing Chl_Atn, Clf_1, and Reg_1 or Clf_2 and Reg_2.

3.4.5 Quantitative Evaluation and Discussion

In Table 3.3, a comparison between the proposed method and several previously reported methods is provided.

It can be seen from the table that the FROC score for the region localization task is 0.841, which is slightly higher than the score reported for the competition winner (Wang *et al.*, 2016) [169] from HMS and MIT by 0.4%. Also, this score is higher than the reference pathologist performance of 0.724 reported in the CAMELYON16 challenge [21].

For the second task, *i.e.*, WSI classification, as mentioned in Section 3.3, WSIs are classified without requiring a separate model to be trained from scratch on several extracted features of the predicted mask (which was done by other methods). Thus, additional computational cost and training time are avoided by preventing the requirement to develop a separate model for WSI classification. As described earlier in Section 3.3.3, for each WSI, the whole-slide probability map is produced by merging the predicted masks of ROIs in sequence. Then, morphological methods (opening and closing operations) are applied to regularize the predicted masks. Finally, each WSI is assigned a score by the average score across all detected regions (average of the probabilities of foreground pixels of the detected WSI mask).

Our AUC score is 0.9703, which is higher than the reference pathologist score and is comparable to the score reported by the competition winner (Wang textitet al., 2016)[169], (as described in Section 3), who developed a 2-stage model and used a dedicated classification network to classify the detected targets. Instead, this approach is an end-to-end model and only uses less than one-third of the images for training.

As shown in the next section, such a design simplifies the network model and speeds up the processing.

The proposed method is additionally evaluated under a stricter WSI-level protocol to mitigate potential patch-level leakage. WSIs are first partitioned, and patches are extracted only after the WSI-level split is defined, preventing correlated patches from the same WSI from appearing in both training and validation data.

For increased statistical robustness, this setting is evaluated using WSI-level k-fold

Table 3.3 : Comparison of results with top-ranking methods on CAMELYON16 [21]. For the proposed method, WSI-level split results are reported as mean ± standard deviation using WSI-level k-fold cross-validation ($k = 5$).

Team	**Architecture**	**AUC (%)**	**FROC (%)**	**Remarks**	**Resolution**	**WSIs**
Pathologist	N/A	96.6	72.4	Diagnosis of a professional pathologist with no time constraint	40X	-
HMS[a] and MIT [169]	GoogLeNet	99.4	80.7	Apply stain normalization; Hard negative data mining; Ensembled two networks	40X	270 WSIs
HMS[b] and MGH [102]	ResNet	97.6	76.0	Apply FCN; Finetuning pretrained network	40X	270 WSIs
CULab[c]	VGG-16	94.0	70.3	Apply FCN; Finetuning pretrained network	40X	270 WSIs
Guo textitet al. (2019) [57]	Inception-v3 and DCNN	96.6	83.5	Combine Inception-v3 and DCNN	40X	270 WSIs
Ours (patch-split)	ResNet50 backbone	**97.03**	**84.1**	Apply pretrained model	10X	80 WSIs
Ours (WSI-level split, k-fold)	ResNet50 backbone	96.9 ± 0.3	83.8 ± 0.6	Apply pretrained model	10X	80 WSIs
Ours (full training set)	ResNet50 backbone	97.71	84.93	Apply pretrained model	10X	270 WSIs

[a] Harvard Medical School and Massachusetts Institute of Technology.

[b] Harvard Medical School and Massachusetts General Hospital.

[c] Chinese University Lab.

cross-validation (here, $k = 5$) on the selected 80 training WSIs. Unless otherwise stated, results are reported as mean ± standard deviation across folds. The test set is identical to the primary experiments, and only the validation protocol differs. Overall, performance

remains stable relative to the patch-split protocol, suggesting that the method performance remains stable under this stricter protocol for strong performance. Table 3.3 summarizes the results for the primary patch-split protocol (single-run reference) and the WSI-level k-fold protocol.

To further examine the effect of training data size when using a reduced training subset, the proposed framework is additionally evaluated using the full CAMELYON16 training cohort (270 WSIs). The model configuration and evaluation pipeline remain unchanged. A WSI-level partitioning strategy is applied, and patches are extracted only after WSIs are split to avoid correlated patch leakage. Approximately 1.5 million 256×256 patches are sampled from the training WSIs for optimization, and the official test set remains identical to the main experiments. The results obtained using the full CAMELYON16 training set (AUC = 97.71%, FROC = 84.93%) are reported in Table 3.3 to provide an additional reference experiment under increased training coverage. The observed improvement compared to subset-based training is attributed to the increased diversity of training samples, which reduces selection bias and supports more stable generalization performance.

While patch-split results are reported as a reference for consistency with prior work [82], the primary conclusions of this chapter are based on the WSI-level k-fold evaluation results, which provide a leakage-resistant and statistically robust assessment.

3.4.6 Computation Time

One of the goals of this study was to design an accurate and computationally efficient end-to-end model for metastasis detection. According to Bejnordi textitet al. [21], a pathologist who has no time constraints needs about 30 h to detect metastases in 129 WSIs of the CAMELYON16 dataset. This means that, on average, an expert pathologist requires 13.95 min to assess one WSI. To examine the computational efficiency of the method, the time performance of this method is compared with other highly ranked models. For this purpose, the required time of processing an ROI with the size of 2816×2816 at a stride of 32 is calculated when using a single GPU. The results indicate that the proposed method requires substantially less processing time than the compared methods.

Table 3.4 : Comparison of the processing time of the proposed method with previously reported methods for processing an ROI and a WSI.

Team	**Architecture**	**ROI's Runtime (minutes)**	**WSI's Runtime (minutes)**
Ours	End-to-end	0.0051	2.04
Pathologist	N/A	0.034	13.95
HMS and MIT [169]	GoogLeNet	0.671	268.40
HMS and MGH [102]	ResNet	0.729	291.60
CULab	VGG-16	0.979	391.60

The proposed model require 0.0051 min (0.306 s) to process the ROI, *i.e.*, considerably faster than the other methods listed in Table 3.4. The average size of a WSI is equal to 400 ROIs of size 2816×2816. As a result, the processing time for a WSI of this average size can be estimated. In Table 3.4, the computational speed of the proposed model compared with other methods is presented. It can be seen that this method only needs 2.04 min to scan a WSI for detecting metastases. This runtime is shorter than the average time reported for manual assessment by an expert pathologist and is also considerably shorter than the runtime reported for the CAMELYON16 competition winner.

3.4.7 Implementation Details

The proposed method is implemented using the PyTorch library and is tested with a single NVIDIA Quadro RTX 6000 Passive 24 GB GPU. The Adam optimizer is used with the learning rates 1e−3 for the regression branches and 1e−4 for the classification

branches. A weight decay of 1e−3 is applied, and the ReLU function is used as the activation function in the inner layers of the framework.

All hyperparameters were selected using the validation set only, with the test set kept strictly isolated throughout model development. No test-set performance was used to guide model selection or parameter adjustment. The reported hyperparameter values correspond to the final configuration used consistently across all experiments in this chapter.

3.4.8 Discussion and Limitation

The proposed approach has achieved competitive performance for the region localization task compared with earlier works. Most importantly, only less than 1/3 of the training images are used and produced competitive results. Decreasing the number of essential annotated WSIs can help reduce the annotation burden because labeling ground truth data in pathology labs is very expensive and time-consuming. Also, WSIs are used at a lower resolution with four times downsampling, *i.e.*, "10X," where the highest resolution, "40X," has been applied in other existing methods. This contributes to reduced computational cost of the model. Although this model produces a detected dense map of only 1/32 of the input's size, the model maintains effective detection performance. At the same time, the processing time per WSI is substantially shorter than the average manual assessment time reported in the CAMELYON16 study.

Our proposed method can detect metastatic regions without considering the costly and time-consuming processes of hard negative and positive sample mining with fewer WSIs and lower-resolution inputs during training. This is because developing a fast and reliable method with fewer WSIs for training while reducing the calculation time and computation resources has been the goal in this work. However, the effects of a small training dataset and the disinclination in using hard sample mining can be observed in some hard samples. Therefore, the model still has limitations. Fig. 3.9 shows an example of a WSI with hard positive pixels. It can be seen that, although this method has detected most of the metastatic regions correctly, some metastatic regions remain undetected. Also, as shown in Fig. 3.10, even though this method has detected most of the regions correctly, there



Figure 3.9 : Visualization of the breast cancer detected with the proposed method. (a) The whole slide image is overlaid with tumor polygons (indicated in blue). (b) The ground truth of tumor regions. (c) The predicted tumor regions by the proposed method.

are minor discrepancies in some spots that are hardly visible in low resolutions. Fig. 3.10 shows the ground truth mask and the detected mask.

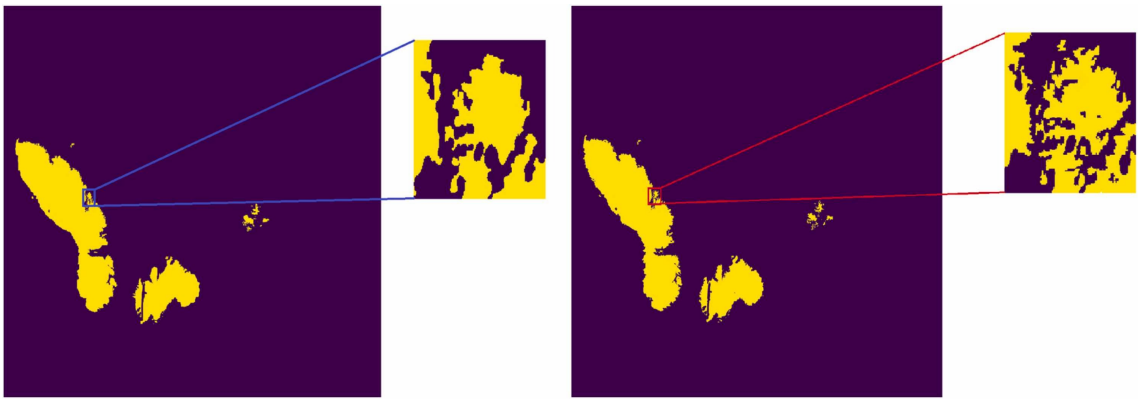


Figure 3.10 : (a) The ground truth mask along with the zoomed-in view of a small area of the tumor mask, and (b) the predicted mask along with the corresponding zoomed-in view of the same area. A slight difference can be observed between the ground truth and the detected mask.

3.5 Summary

This chapter examined several limitations of metastasis detection and localization methods in breast cancer histopathology. These limitations include high computational

demands, dependency on large annotated datasets and difficulty processing whole slide images at high magnifications. To address these issues, a weakly supervised end-to-end CAD framework was proposed, designed for efficient processing at 10× magnification and requiring less than one-third of the annotated data.

This design reduces computational requirements while maintaining competitive detection performance.

The model combines residual connections, regression, and classification tasks, as well as spatial and channel attention modules to perform WSI-level classification and metastasis localization simultaneously. Experimental results on the CAMELYON16 dataset indicate that the proposed method achieves competitive performance in both metastasis localization and WSI classification. The system has demonstrated the ability to process a WSI in just 2.04 minutes on a single GPU, which is shorter than the average time reported for manual assessment by an expert pathologist and shorter than the runtime reported for several previously published automated methods. These results suggest that the proposed framework provides an efficient approach for metastasis detection while reducing computational requirements and annotation demands.

Despite these encouraging results, several limitations remain. The use of 10× magnification improves computational efficiency but may limit sensitivity to fine-grained cellular details, particularly for small metastatic regions. In particular, improving spatial context modeling and localization precision remains important for further advancing histopathology-based metastasis detection. Improved spatial context modeling can enable the model to better capture relationships between distant tissue regions within WSIs, leading to a more comprehensive representation of histopathological structures. Additionally, improving the delineation of tumor boundaries remains important, particularly for small or irregular metastatic regions in complex and heterogeneous tissue environments.

Addressing these challenges may further improve spatial localization performance and strengthen the robustness of metastasis detection in whole-slide histopathology images.

These limitations motivate the methodological refinements introduced in the following

chapter. In particular, Chapter 4 presents an enhanced weakly supervised framework designed to improve contextual feature modeling and boundary localization accuracy while maintaining practical computational feasibility and reduced annotation requirements.

Chapter 4

A Weakly-Supervised Data-Efficient Framework for Metastasis Detection and Localization in Breast Histopathology Images

While the end-to-end metastasis detection approach proposed in Chapter 3 demonstrates substantial improvements in computational efficiency and accuracy by utilizing lower-resolution WSIs (10×) and a reduced dataset, certain limitations persist. Challenges such as insufficient context-awareness, suboptimal boundary delineation, and opportunities for further optimization in training data utilization still affect its performance. In particular, the absence of a decoder in the Chapter 3 framework limits its ability to recover high-resolution spatial details, which constrains boundary refinement and precise localization of metastatic regions. As a result, accurately capturing spatial relationships and refining tumor boundaries remain key challenges.

To address these limitations, this chapter introduces an enhanced weakly supervised framework for metastasis detection and localization in breast histopathology WSIs. Building upon a ResNet backbone, the proposed architecture enhances feature extraction and spatial context modeling to improve metastasis detection and localization in WSIs.

This formulation goes beyond coarse lesion-level localization by producing spatially refined prediction maps that enable more precise, boundary-aware localization of metastatic regions.

Although the task is referred to as localization, the proposed framework produces dense prediction maps that approximate boundary delineation of metastatic regions. This capability is referred to as *boundary-aware localization*, which bridges the gap between conventional lesion-level localization (typically evaluated by FROC) and pixel-wise seg-

mentation (evaluated by Dice and IoU).

As a weakly supervised approach, this new approach continues to operate on lower-resolution WSIs at 10× while further reducing computational costs and annotation requirements. Unlike the previous method, which used one-third of the annotated CAMELYON16 dataset, this work improves data efficiency by training on only one-sixth of the annotated dataset (40 WSIs versus 80 WSIs). Despite the reduced training data, the model maintains competitive performance through several architectural enhancements.

To address the limited spatial context modeling of conventional CNNs, GRUs [33] are incorporated to capture dependencies across distant tissue regions and improve contextual consistency across tissue structures. To address the limited multi-scale representation capability of conventional CNNs, the ASPP [31] module is integrated to aggregate features at multiple receptive fields while preserving fine-grained spatial details. Additionally, to mitigate the insufficient attention to tumor-relevant features in conventional CNNs, CBAM [173] is incorporated to guide the network toward tumor-relevant spatial and channel features while suppressing background responses.

Finally, to compensate for the loss of high-resolution spatial details caused by repeated downsampling in CNN architectures, skip connections [132] are employed to transfer high-resolution spatial features from the encoder to the decoder. This helps preserve spatial information during feature reconstruction and supports more precise metastasis localization, particularly for small or fragmented tumor regions.

Collectively, these enhancements address the gaps identified in the previous approach by: 1) enhancing context awareness to capture spatial relationships within WSIs more effectively, 2) improving the boundary precision of the detected tumors for sharper tumor edge delineation, and 3) enabling more efficient training with minimal annotated data.

Experimental results demonstrate strong performance, indicating the model's ability to capture spatial and contextual patterns in WSIs. Compared to the previous approach, this framework achieves higher accuracy in both detection and localization, while using only half the training data, highlighting its improved data efficiency.

4.1 Introduction

Metastasis detection and localization in WSIs remain challenging due to the inherent complexity of tissue structures and imaging variability. One key challenge is the need for robust contextual understanding, as metastatic regions vary significantly in size, shape, and spatial distribution within the tissue. Accurately detecting these regions requires the ability to model global dependencies and capture intricate spatial relationships.

Another critical obstacle lies in preserving high-resolution spatial details. Tumor boundaries are often irregular or fragmented, making it difficult to distinguish between metastatic and healthy regions without precise boundary refinement. This issue is compounded by the need to detect small or sparsely distributed metastases, which are easily missed during standard downsampling processes.

Scale variability further complicates the detection process, as metastatic regions can exhibit different levels of visual complexity within the same WSI. This variation arises due to differences in tissue heterogeneity, staining inconsistencies, and morphological diversity of tumor cells. Such complexity makes it challenging to distinguish between benign and malignant areas, particularly for tiny metastases with subtle visual cues. Effective multi-scale analysis is essential to capture both micro- and macro-level tumor features. Additionally, ensuring the model focuses on the most relevant spatial and channel features is vital to minimize false positives and false negatives, especially in noisy or complex tissue environments.

To address the complexities of metastasis detection and localization in WSIs, a robust, weakly supervised, and data-efficient algorithm is proposed for metastasis detection and localization. The approach is designed to overcome the limitations of existing methods by integrating advanced computational modules. These modules work synergistically to enhance both classification and localization performance. Collectively, these components address key limitations of conventional CNN architectures, including restricted spatial context modeling, limited multi-scale representation, and insufficient attention to tumor-relevant features.

To capture the intricate spatial dependencies inherent in WSIs, GRUs [33] are incorporated into the residual layers of the encoder. This integration enables the model to learn relationships between distant regions. As a result, it enriches contextual understanding and provides a more comprehensive representation of histopathological features. This global dependency modeling is essential for accurately detecting metastases across diverse tissue structures.

To address scale variability in histopathology images, the ASPP module [31] is employed in the bottleneck layer to aggregate features at multiple receptive fields. By processing features at varying receptive fields, ASPP [31] allows the model to effectively identify metastases of different sizes and shapes. This capability addresses the challenge of scale variability. It ensures robust detection across both micro- and macro-level tumor features.

To improve feature selection in complex tissue environments, CBAM [173] is incorporated to guide the network toward tumor-relevant spatial and channel features. Applied in both the bottleneck and skip connections [132], CBAM [173] directs the model's focus toward tumor-relevant spatial and channel features. This targeted attention minimizes false positives and negatives. It is particularly effective in noisy and complex tissue environments, improving the reliability of the system.

Unlike the framework presented in Chapter 3, which prioritizes computational efficiency through a lightweight architecture without an explicit decoder stage, the framework in this chapter introduces an encoder–decoder structure with skip connections to better preserve spatial information during feature reconstruction. To preserve fine-grained spatial details crucial for accurate localization, skip connections [132] bridge the encoder and decoder. This design enables the reconstruction of higher-resolution, spatially detailed localization maps while retaining contextual features extracted by the encoder.

By transferring high-resolution features directly to the decoder, skip connections [132] enable the model to maintain critical spatial details. This is especially beneficial for detecting small or fragmented metastatic regions. These regions are often missed during conventional downsampling processes.

By combining these modules, this framework offers a comprehensive solution to the challenges of metastasis detection and localization. Importantly, it achieves this while using only 1/6 of the annotated CAMELYON16 dataset at a reduced 10× resolution. the model remains computationally feasible for WSI analysis while improving spatial precision and localization quality compared with the previous framework. Furthermore, instead of relying on pixel-level annotations and using ground-truth masks for input patches, a dual-task learning framework is adopted that leverages patch-level classification and regression. The classification branch determines whether a patch contains metastatic tissue, while the regression branch estimates the tumor ratio within the patch, providing a quantitative measure of metastatic burden. This approach enables metastasis detection and localization on WSIs without requiring pixel annotation masks. It streamlines the annotation process while maintaining high accuracy in identifying and quantifying metastatic regions. The contributions made in this chapter are listed below:

- A highly efficient framework that simultaneously detects and localizes metastasis from breast histopathology images is proposed without the need for pixel-level annotation during training.
- The model integrates GRUs [33], ASPP [31], CBAM [173], and skip connections [132] to enhance feature extraction, spatial context modeling, and localization accuracy, achieving superior localization precision, particularly for small metastases and tumors with complex boundaries.

Experimental results indicate that the proposed framework improves boundary refinement and supports the detection of metastases with irregular shapes and varying spatial scales, contributing to more precise localization results.

Despite its increased complexity compared to the previous approach and the inclusion of a decoder to preserve high-resolution output, the model remains computationally efficient. It processes WSIs faster than many state-of-the-art methods while ensuring superior accuracy and reliability.

The remainder of this chapter is organized as follows: Section 4.2 presents a comprehensive review of the recent works related to the techniques used in this approach. Section 4.3 provides the details of the proposed model and its architectural framework, emphasizing key components and contributions. Section 4.4 discusses the implementation and evaluation of the model. The experimental results and their significance are analyzed in Section 4.5. Section 4.6 explores the impact of individual components and design choices. The computational efficiency of the model is assessed in Section 4.7. Finally, Section 4.8 synthesizes the findings, highlights the significance of the results, and outlines potential future directions.

4.2 Related Work

Advancements in CAD systems for metastasis detection have tackled significant challenges in histopathological image analysis [169, 102, 98, 99, 63]. These challenges include robust feature extraction, contextual understanding, computational efficiency, and data limitations, e.g., [102, 98, 187, 77]. Researchers have developed diverse approaches, such as feature ensemble methods, attention mechanisms, multi-scale analysis, and weakly supervised learning, to enhance the accuracy and reliability of CAD systems, e.g., [99, 96, 139, 185, 142]. Some representative approaches developed for breast cancer detection are summarized below.

4.2.1 Feature Ensemble Methods

Feature ensemble methods, e.g., [78, 60, 170, 11] integrate multiple models or feature extraction techniques to enhance performance. They improve analysis and interpretation by leveraging diverse architectures, such as CNNs, e.g., [78] and transformers, e.g., [170, 11]. Combining features from different scales, modalities, or representations strengthens model robustness and reduces bias. These methods e.g., [78, 60, 170] enhance generalization across datasets, addressing variations in tissue structures and imaging inconsistencies.

However, ensemble models face several challenges. Their high computational com-

plexity and memory demands make them less feasible for real-time deployment, especially in large-scale WSI analysis. Additionally, ensemble methods risk overfitting, particularly when trained on limited datasets and often lack interpretability, making clinical adoption difficult. While they enhance the overall detection performance, they do not naturally improve the accuracy of tumor boundary identification, highlighting the need for spatially-aware architectures to achieve more precise localization [114].

To improve classification robustness across datasets, Kassani *et al.* [78] proposed an ensemble approach combining VGG19, MobileNet, and DenseNet architectures. This design leverages the strengths of multiple architectures to capture diverse features, enhancing context sensitivity and model robustness. Their work highlights the potential of feature ensembles in addressing the variability in histopathological data. However, it also points to challenges in generalizing across datasets with varying sizes and properties. Additionally, the ensemble approach significantly increases computational complexity, making real-time clinical applications difficult. The lack of explicit spatial attention mechanisms may also limit its ability to refine tumor boundaries effectively. Guo *et al.* [57] introduced v3_DCNN, which combined Inception-v3 for tumor pre-selection with a dense convolutional neural network for segmentation. This approach significantly reduced processing time, making it suitable for high-throughput scenarios.

Similarly, Singha *et al.* [145] introduced AlexSegNet, which fused feature maps within the encoder and employed skip connections [132] in the decoder for improved nucleus segmentation. In another approach, Ashurov *et al.* [11] combined attention mechanisms with transfer learning architectures like Xception, ResNet50, and DenseNet121 to enhance breast cancer detection across different magnifications. This integration demonstrated that leveraging diverse architectures enables more comprehensive feature extraction, contributing to better detection performance. However, the reliance on pre-trained models trained on natural images may restrict adaptability to histopathological textures, requiring additional domain adaptation strategies.

4.2.2 Attention Mechanisms in Histopathology Image Analysis

Attention mechanisms in deep learning selectively emphasize the most relevant features in an input while suppressing less important information, enhancing model performance in tasks such as classification, localization, and segmentation. They dynamically allocate computational focus, mimicking human perception by directing attention to key areas rather than processing the entire image uniformly [24, 118]. Attention can be applied across spatial dimensions, where it highlights important regions in an image, or across feature channels, where it prioritizes meaningful patterns in learned representations. In image analysis, attention mechanisms enhance localization, boundary refinement, and noise suppression [119, 177]. These capabilities make them particularly valuable in complex tasks like medical image analysis, where distinguishing subtle patterns is crucial for accurate diagnosis.

Advancements in attention mechanisms have significantly enhanced feature selection in medical image analysis, particularly for tumor localization and classification [119, 138, 177]. By dynamically adjusting attention weights on input features, mechanisms such as CBAM [173] or attention gates [119], improve feature relevance.Their strength lies in their ability to emphasize spatially and channel-wise informative features, thereby refining deep learning models in histopathological image analysis.

While attention mechanisms effectively enhance feature selection and improve model focus on diagnostically relevant regions, certain types, such as channel-wise and self-attention mechanisms, may not explicitly refine spatial details [69, 162, 173]. These methods primarily emphasize feature importance across channels or global contexts but do not inherently improve fine-grained boundary delineation, which is critical in tasks like tumor segmentation. Consequently, without additional spatially-aware components, they may struggle to precisely localize tumor margins, particularly in cases of diffuse or irregular metastases [69, 162, 138]. Accurately detecting and localizing metastases in WSIs requires a holistic approach that ensures both precise boundary delineation and robust feature representation. While focusing on diagnostically relevant regions is essential, pre-

serving spatial coherence helps maintain the continuity of tissue structures, preventing fragmented or inconsistent segmentation. Additionally, capturing scale variations allows the model to recognize metastases of different sizes and morphological patterns, which is crucial for distinguishing between small micrometastases and larger lesions. Finally, efficient processing of high-resolution histopathological data ensures that fine-grained spatial details are retained without excessive computational overhead, enabling reliable and scalable metastasis detection.

Recent research has explored the use of attention mechanisms in histopathology image analysis. Yao *et al.* [181] introduced DeepAttnMISL, an attention-based MIL framework for survival prediction using WSIs. Unlike traditional models that rely on ROI selection or handcrafted feature aggregation, their approach integrates Siamese Multiple Instance Fully Convolutional Networks (MI-FCN) with attention-based MIL pooling. This design adaptively highlights critical regions within WSIs, improving feature selection and model interpretability. However, it lacks spatial consistency and multi-scale feature extraction, limiting its ability to refine tumor boundaries or detect metastases of varying sizes.

Wang *et al.* [170] introduced Semantically Relevant Contrastive Learning (SRCL), a self-supervised learning strategy designed to enhance histopathological image analysis by leveraging instance-level feature alignment. Their framework employs CTransPath, a hybrid CNN-Transformer backbone integrating local feature extraction via CNNs and global context modeling via Swin Transformers. The multi-head self-attention mechanism in Swin Transformers enables effective feature representation across multiple scales, making the model highly adaptable to diverse histopathological tasks, including patch classification, mitosis detection, and weakly supervised WSI classification. Nevertheless, contrastive learning relies heavily on feature similarity but does not inherently enforce spatial coherence, which is crucial for delineating complex metastases. Furthermore, the computational burden of pretraining remains a challenge, necessitating optimization strategies to ensure efficiency in real-world clinical applications. Domain-specific innovations have also advanced CAD systems' capabilities. Histoseg by Wazir *et al.* [171] used attention

mechanisms for cell segmentation, combining global and local feature representations with a multi-loss function.

In another approach, Qiu *et al.* [127] introduced a hierarchical multimodal fusion framework that leverages multi-level attention mechanisms to integrate pathology images and genomic data for cancer classification. The AHM-Fusion module applies attention-based guidance at both early and late fusion stages, allowing the model to prioritize critical features and improve modality interaction. Additionally, the PathIn-NL model employs weakly supervised noisy label learning to extract patch-level features from WSIs without requiring pixel-level annotations. This approach demonstrates strong classification performance on glioma and lung cancer datasets, showcasing the potential of attention-driven multimodal fusion in medical imaging. The framework prioritizes feature weighting and multimodal interactions but lacks multi-scale feature extraction, limiting its ability to detect metastases of varying sizes and complex structures. This is due to its focus on attention-based feature selection rather than explicit spatial scaling techniques, such as ASPP [31] or multi-resolution analysis. As a result, it struggles with capturing tumors across different scales, impacting the performance in heterogeneous tissue environments.

Similarly, Shi et al. [142] introduced CSAM-ResNet, integrating Channel and Spatial Attention Mechanisms (CSAM) with residual learning. This enhanced the model's ability to capture global and local features, boosting diagnostic accuracy. Despite its advantages, CSAM-ResNet faces certain limitations. While it enhances feature selection, it lacks an explicit mechanism for handling intra-tumoral heterogeneity, where varying tumor compositions within the same slide may require more adaptive learning. Additionally, its reliance on attention mechanisms without explicit multi-scale feature extraction may limit its performance when detecting metastases of widely varying sizes.

4.2.3 Multi-scale Analysis

Multi-scale analysis has become a crucial approach in histopathology image analysis, addressing the challenge of capturing both fine-grained cellular details and broader tissue architecture. By integrating features across multiple resolutions, multi-scale methods en-

hance metastasis detection, tumor classification, and localization accuracy [161, 94]. This technique enables models to detect patterns that vary in size and structure more effectively. This is particularly important in WSIs, where metastatic regions may appear at different scales and levels of detail.

Despite these advantages, existing multi-scale methods, e.g., [161, 94, 139] often struggle with computational efficiency and spatial coherence. These approaches, e.g., [161, 139] rely on high-resolution images, increasing processing time and memory requirements. Additionally, while these methods [161, 94, 139] improve feature representation, they do not always refine tumor boundary delineation or adapt well to varying tissue microenvironments. These limitations highlight the need for a more efficient, structured, and adaptive approach to multi-scale learning in metastasis detection.

ScanNet by Lin *et al.* [98] employed a fully convolutional divide-and-conquer strategy, aiming to process WSIs at different scales without exhaustive patch-wise analysis. While this method optimized large-scale image processing, it still struggled with contextual feature integration across different scales. The framework lacked a mechanism for dynamically adjusting receptive fields, making it less adaptable to tumors of varying morphologies. To further optimize speed and efficiency, Fast ScanNet [99] refined this approach by introducing anchor layers and asynchronous sample prefetching, reducing latency in large-scale WSI scanning. However, this optimization came at the cost of feature resolution tradeoffs, as the model prioritized processing speed over fine-grained feature preservation. This limitation particularly impacted boundary delineation, making it less reliable for precise tumor localization, especially in cases with diffuse or irregular metastases.

Building on multi-scale efficiency, Zhang *et al.* [187] introduced a three-stage framework integrating patch-based classification, tumor localization, and WSI-level classification. The model leveraged CycleGAN for color normalization and a hybrid DPN68-A with Swin-Transformer architecture. This multi-scale strategy allowed the model to capture diverse tumor features at different resolutions, effectively improving classification robustness. However, while the approach demonstrated strong feature extraction, it relied on

separate stages for classification and localization, increasing computational overhead and complexity. Additionally, managing architectural complexity and computational demands remained a challenge, particularly when scaling to large WSIs.

4.2.4 Weakly Supervised Learning

Weakly supervised methods in histopathology images leverage high-level labels, such as WSI-level annotations, to train models without requiring pixel-wise localization. These approaches reduce the burden of manual annotation while effectively learning tumor localization and classification. However, challenges remain in capturing fine-grained spatial details and improving boundary delineation, necessitating advanced attention mechanisms and multi-scale feature extraction for enhanced precision.

Efficiency and scalability remain vital concerns for CAD systems.

Pioneering this direction, Li *et al.* [92] introduced DSMIL, a dual-stream MIL framework designed for WSI classification and metastasis detection using only slide-level annotations.The model combines self-supervised contrastive learning with a multi-scale pyramidal fusion approach to enhance patch-level representations. The dual-stream aggregator effectively models instance relationships, enabling improved identification and localization of tumor regions without requiring dense annotations. The method demonstrated strong classification and localization performance across multiple datasets and improved upon earlier MIL-based approaches. While DSMIL advances weakly supervised detection by integrating multi-scale features and contrastive learning, its reliance on memory-intensive bag construction may limit scalability. Future work could investigate more efficient aggregation strategies or adaptive patch sampling to support broader application in large-scale WSI analysis.

Extending this direction, Xia *et al.* [174] proposed a weakly supervised segmentation method using point annotations. This method relied on dual-task learning to enhance segmentation accuracy. These domain-specific methods highlight the adaptability of CAD systems to various histopathological tasks but emphasize the need for scalable solutions to

handle large, diverse datasets effectively. Incorporating attention, Li *et al.* [96] proposed SA-MIL, integrating self-attention modules into a MIL framework to enhance segmentation without full supervision. While effective, SA-MIL encountered difficulties in achieving precise boundary delineation compared to fully supervised methods.

Combining learning paradigms, Xie *et al.* [176] developed MoCo-CLAM, a framework combining contrastive learning with clustering-constrained attention mechanisms under a weakly supervised MIL setting. This framework adapted features to coarse labels, improving data utilization. However, it faced challenges in tasks requiring finer-grained classification. Comparatively, Kang *et al.* [77] compared Fully supervised learning (FSL) and weakly supervised learning based on multiple instance learning (WSLMIL) for lymph node metastasis detection, evaluating backbone networks and aggregation methods to optimize classification. Valieris *et al.* [160] proposed a self-supervised, attention-based weakly supervised model to classify HER2 status from histopathological images. Finally, Gao *et al.* [49] introduce HTransMIL, a transformer-based MIL model, using only slide-level labels for WSI breast cancer classification. The model captures global context and visualizing deep features for interoperability.

4.2.5 Motivations

Collectively, the above surveyed prior works represent significant advancements in CAD systems for metastasis detection and localization. Approaches such as feature ensembles, attention mechanisms, multi-scale analysis, and weakly supervised learning address critical challenges in histopathological image analysis.

However, they also reveal ongoing limitations that motivate this approach. Attention mechanisms enhance feature selection but often lack spatial consistency [181, 170, 127], leading to imprecise tumor boundary delineation. Multi-scale analysis enables the capture of tumor features across different scales; however, various models (e.g., Lin *et al.* [98, 99]; Zhang *et al.* [187]) rely on localized region-based processing, which can disrupt spatial coherence and increase computational complexity. Weakly supervised learning reduces the dependency on dense annotations but often struggles with precise localization due to

the absence of structured spatial refinement. Additionally, many models fail to effectively integrate global context with fine-grained details, limiting their adaptability to metastases of varying sizes.

These limitations motivate the development of more robust and scalable CAD frameworks, such as the one proposed in this chapter. In this chapter, a robust and efficient CAD framework proposed, that combines GRUs [33] for spatial context modeling, ASPP [31] for multi-scale feature extraction, CBAM [173] for attention-based feature refinement, and skip connections [132] for preserving spatial details during feature reconstruction. This integration addresses key challenges in metastasis detection and localization.

Unlike previous models, which rely on either attention, multi-scale, or weak supervision alone, the proposed approach integrates these techniques into a unified, efficient framework. Specifically, the proposed model incorporates global dependency modeling and multi-scale feature aggregation to enhance its understanding of tissue variability. Targeted attention mechanisms and fine-detail preservation further improve accuracy and reliability. These advancements are achieved while maintaining practical computational feasibility and reduced data requirements.

4.3 The Architecture of the Proposed Model

The proposed model is an end-to-end weakly supervised architecture designed for simultaneous metastasis detection and localization on WSIs. The model is built upon the ResNet50 backbone and integrates advanced modules such as Bidirectional GRUs (Bi-GRUs) [33], ASPP [31], CBAM [173], and skip connections [132].

As illustrated in Fig. 4.1, the model first processes image patches of size 256×256 extracted from WSIs, which are passed through a ResNet-based encoder. To enhance spatial coherence and address the limited spatial context modeling of conventional CNNs Bi-GRUs [33] are integrated at each ResNet layer (see the detailed architecture of the proposed model in Fig. 4.2). Then, the extracted features are further enhanced by integrating multi-scale refinement via ASPP [31] at the bottleneck, which addresses the limited multi-

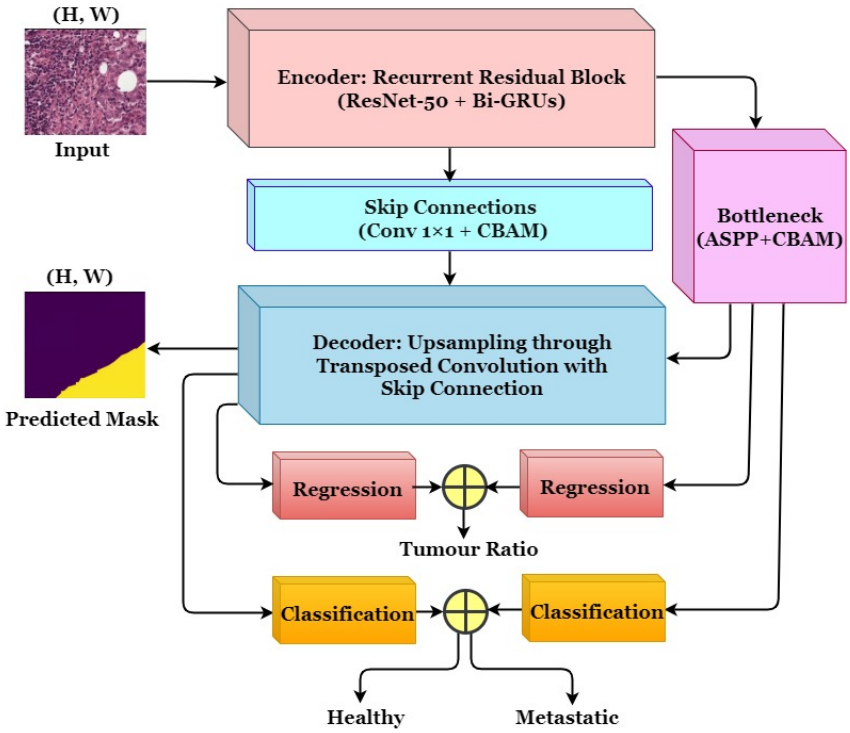


Figure 4.1 : Overview of the proposed architecture for metastasis detection and localization in histopathology images. Input patches are processed by a ResNet-50 + Bi-GRU encoder. Multi-scale contextual features are refined at the bottleneck using ASPP and CBAM. The decoder reconstructs dense localization maps via transposed convolutions with CBAM-enhanced skip connections. Dual-task learning is performed through classification and tumor ratio regression at both the bottleneck (global features) and decoder (high-resolution features). The outputs from the bottleneck and decoder branches are combined via averaging to produce the final classification and regression predictions, enabling joint metastasis classification and spatially resolved localization.

scale representation capability of conventional CNNs by extracting multi-scale features against variations in tumor size and shape. Following ASPP [31], CBAM [173] enhances spatial and channel attention, amplifying diagnostically relevant regions while suppressing irrelevant background information to ensure that critical tumor regions are prioritized. In

the decoder, CBAM-enhanced skip connections [132] facilitate the reconstruction of precise localization maps. Additionally, classification and regression tasks at both the bottleneck (global feature level) and decoder (high-resolution feature level) guide metastasis detection and localization without requiring pixel-level annotations during training.

Specifically, the classification branch determines the presence of metastases within an image patch, while the regression branch estimates the tumor ratio, providing a quantitative measure of metastatic burden. These tasks are applied at two key stages: the bottleneck layer, where global contextual features are extracted, and the final decoder layer, which refines high-resolution spatial details. The outputs from both levels are then combined to balance global and fine-grained predictions, supporting more accurate localization and improving overall prediction quality.

The decoder reconstructs high-resolution localization maps through progressive upsampling, transposed convolutions, and CBAM [173] refined skip connections [132]. At each stage, upsampled decoder features are merged with corresponding encoder outputs that have been refined by CBAM, preserving high-resolution spatial details while improving boundary delineation. The final transposed convolution restores the feature map to the original input resolution $(1, H, W)$, ensuring precise localization.

Thus, by leveraging GRUs [33] for contextual consistency, ASPP [31] for multi-scale feature extraction, and CBAM [173] for attention refinement, the model improves spatial precision by explicitly addressing key limitations of conventional CNNs, including restricted spatial context modeling, limited multi-scale representation, and insufficient attention to diagnostically relevant features and localization quality while remaining computationally feasible for WSI analysis. This design also reduces annotation requirements by avoiding pixel-level supervision during training. In addition, unlike traditional models that rely on high-magnification ($40\times$) WSIs and extensive annotations, this approach operates effectively at $10\times$ magnification while utilizing only a fraction of annotated data.

The detailed architecture of the proposed model is illustrated in Fig. 4.2, and is described step-by-step below, with each stage building upon the outputs of the preceding

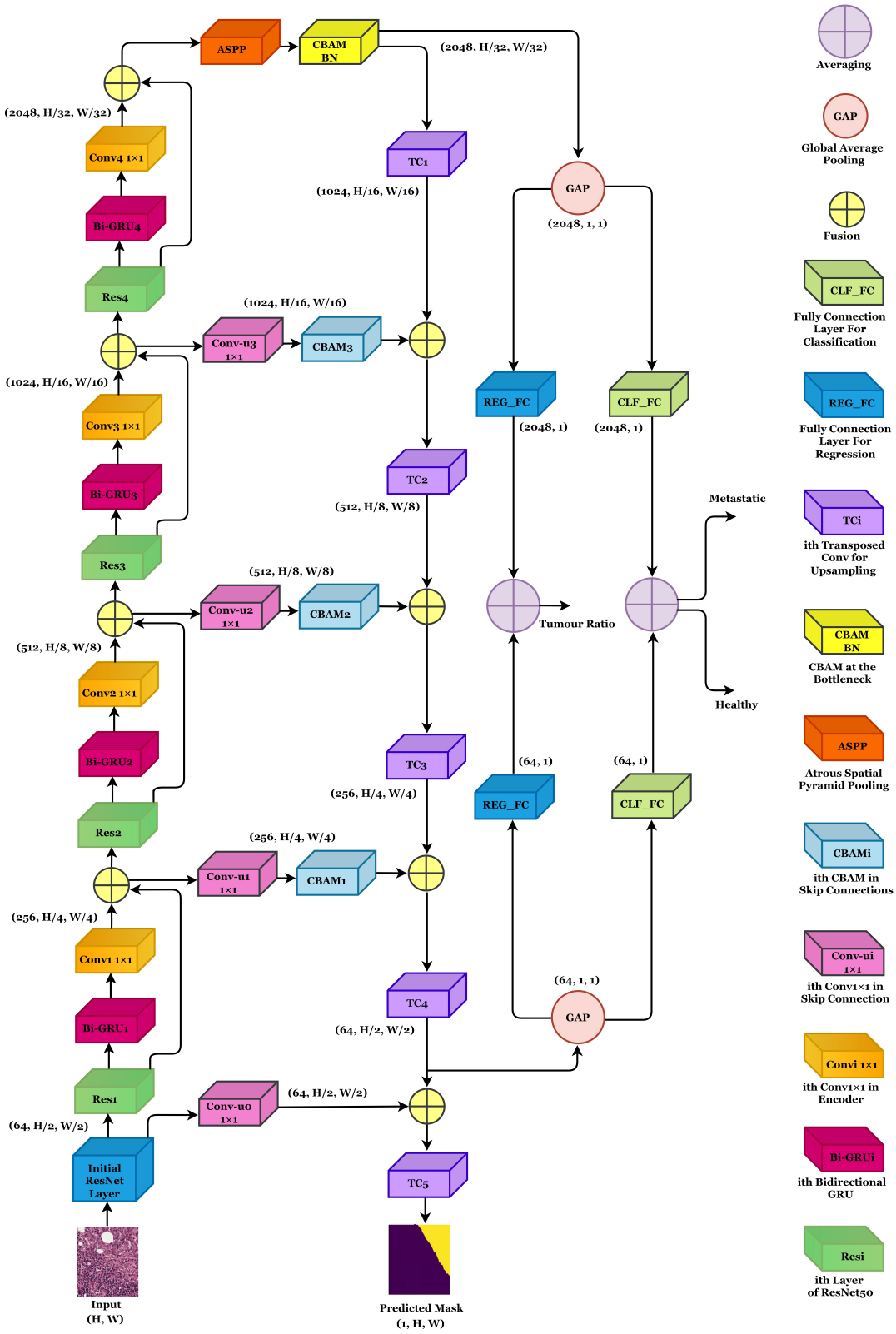


Figure 4.2 : The detailed framework of the proposed model.

steps to ensure seamless integration and optimized performance.

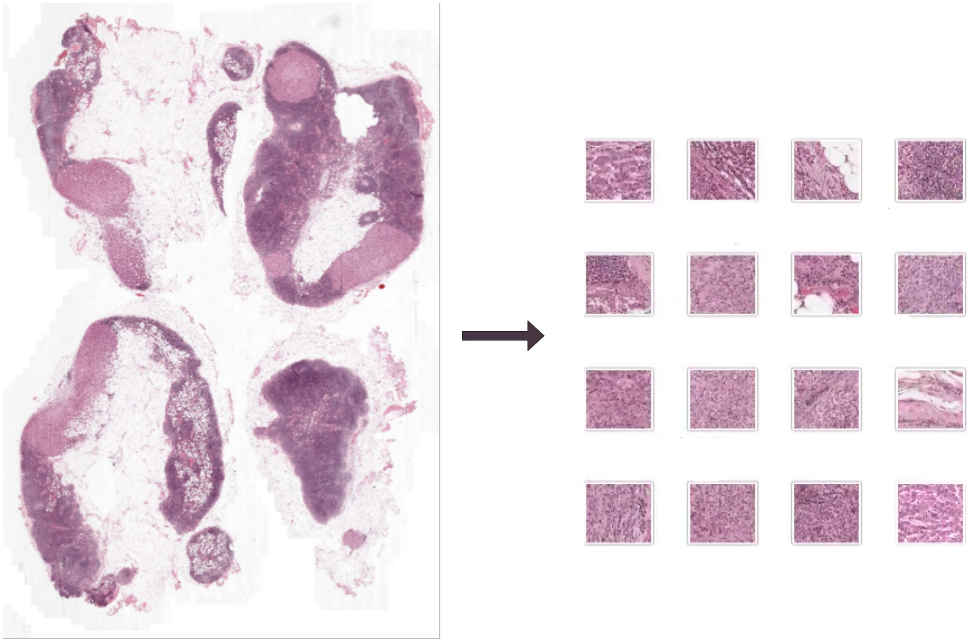


Figure 4.3 : An example of a WSI (left) and several extracted patches from it (right).

4.3.1 Input Preprocessing and Initial Feature Extraction

The CAMELYON16 WSIs used in this chapter are first partitioned at the WSI level before patch extraction. From the selected training WSIs, a total of 600,000 patches of size 256×256 are extracted for model optimization.* Among these, 300,000 patches contain metastatic tissue, while the remaining 300,000 represent non-metastatic regions, ensuring a balanced dataset for training.

Each patch is labeled with a binary class (healthy or metastatic) along with a tumor ratio that quantifies the proportion of metastatic content within the patch. The tumor ratio serves as an auxiliary regression target to guide the training strategy, as further discussed in Section 4.3.5. Although pixel-level annotations assist in generating these labels and tumor ratios during data preparation, the model training process itself does not directly rely on pixel-wise supervision. Intensity normalization is applied to standardize

*The model is designed to adapt seamlessly to varying input dimensions.

input values, improving feature consistency across samples. During the prediction phase, larger patches of 2048×2048 with an overlap size of 128 pixels are used to accelerate the processing of WSIs. Fig. 4.3 shows an example of a WSI and several patches extracted from the WSI image.

As depicted in the initial stage of Fig. 4.2 (shown in blue), patches are passed through the initial layers of ResNet-50. This stage produces an initial feature map, denoted as F_0, with dimensions $(64, H/2, W/2)$, where H and W represent the height and width of the input patch, respectively.

The ResNet-50 backbone extracts hierarchical features through its residual layers. These layers progressively reduce the spatial resolution of the feature maps while increasing their depth, enabling the model to capture a range of feature abstractions. These hierarchical features serve as the foundation for further refinement and processing in the subsequent modules of the architecture.

4.3.2 Integration of GRUs with ResNet

WSIs are vast, high-resolution images with intricate tissue structures and spatial dependencies that are essential for accurate analysis. Tumor regions often exhibit complex morphological patterns and contextual relationships that extend beyond localized patches, requiring a mechanism to capture these dependencies across spatial dimensions. CNNs process images in a spatially independent manner, meaning they analyze local features in patches but fail to capture the broader context of tissue structures. This is particularly problematic in metastasis detection, where tumor regions exhibit complex spatial dependencies and morphological variations that extend beyond isolated patches. To capture temporal dependencies in sequential data, Recurrent Neural Networks (RNNs) have been widely used. However, traditional RNNs suffer from vanishing gradient issues, making it difficult to retain long-range dependencies [66, 22]. To address these issues, GRUs [33] were introduced as an optimized variation of RNNs. GRUs [33] incorporate two gating mechanisms—the reset gate and the update gate—which regulate the flow of information by selectively retaining important features and discarding irrelevant ones. This allows the

model to efficiently learn long-term dependencies while maintaining computational efficiency, making it particularly effective for deep learning tasks requiring sequential context retention.

In this work, instead of processing text or time-series data, GRUs [33] sequentially process spatial regions of feature maps, allowing the model to establish dependencies between adjacent tissue regions. This spatial recurrence ensures that the extracted features preserve contextual continuity, improving tumor detection and boundary delineation.

In addition, unlike standard GRUs, which process information in a single direction, the proposed approach integrates Bi-GRUs to propagate information both forward and backward across spatial dimensions.

As illustrated in the top row of Fig. 4.2, feature maps produced by ResNet layers are passed through corresponding Bi-GRUs [33] modules to refine spatial dependencies and enrich feature representations.

Let $\mathbf{F}_{\text{ResNet}_i}$ denote the output of the i-th ResNet block, where $i = 1, 2, 3, 4$. Each $\mathbf{F}_{\text{ResNet}_i}$ is passed through a Bi-GRU module to obtain a context-aware representation $\mathbf{F}_{\text{GRU}_i}$, defined as:

$$\mathbf{F}_{\text{GRU}_i} = \text{BiGRU}_i(\mathbf{F}_{\text{ResNet}_i}) \tag{4.1}$$

Shallower ResNet layers generate local representations such as edges and textures, which are refined by the corresponding Bi-GRU modules, i.e., $\mathbf{F}_{\text{GRU}_1}$ captures enriched low-level patterns based on $\mathbf{F}_{\text{ResNet}_1}$. Deeper layers, including $\mathbf{F}_{\text{GRU}_2}$, $\mathbf{F}_{\text{GRU}_3}$, and $\mathbf{F}_{\text{GRU}_4}$, progressively encode mid-level features, high-level semantic structures, and global contextual cues, respectively.

Since GRUs [33] operate on sequential data, the spatial dimensions of feature maps are reshaped into sequences. As illustrated in Fig. 4.4, each feature map $\mathbf{F}_{\text{ResNet}} \in \mathbb{R}^{C \times H' \times W'}$ —where C is the number of channels, and H', W' are the height and width of the feature map—is reshaped into a sequence $\mathbf{P} \in \mathbb{R}^{N \times C}$, where $N = H' \times W'$. Each vector $\mathbf{P}_t \in \mathbb{R}^C$ corresponds to the feature at spatial position t.

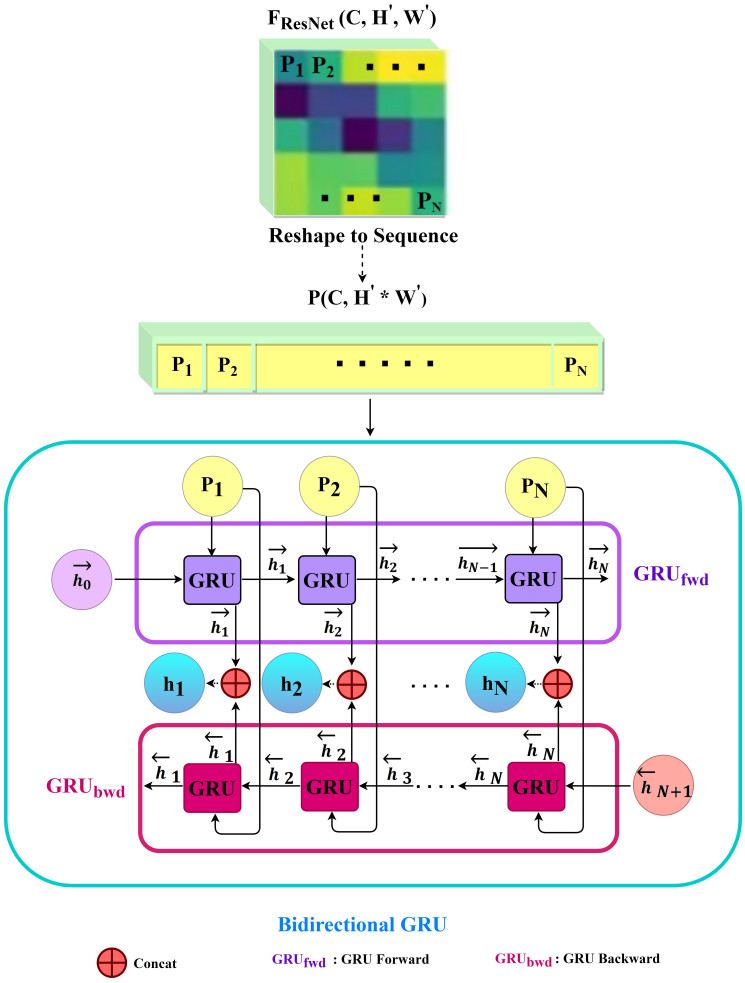


Figure 4.4 : Illustration of Bi-GRUs embedded in ResNet for feature refinement. Feature maps from each ResNet layer are reshaped into sequences and processed by Bi-GRUs to capture spatial dependencies.

Each Bi-GRU module processes the resulting sequence in two directional passes. The forward GRU (GRU_{fwd}), shown in purple in Fig. 4.4, propagates information from the top-left to bottom-right of the feature map, starting from the initial hidden state h_0. The forward pass is defined as:

$$\overrightarrow{h}_t = \text{GRU}_{\text{fwd}}(\mathbf{P}_t, \overrightarrow{h}_{t-1}), \quad t = 1, \ldots, N \tag{4.2}$$

It generates forward hidden states $\overrightarrow{h}_1, \overrightarrow{h}_2, \ldots, \overrightarrow{h}_N$ by sequentially incorporating information from the previous hidden state $\overrightarrow{h}_{t-1}$ and the current input $\mathbf{P}_t$ $(t = 1, \ldots, N)$,

where each $\mathbf{P}_t$ represents the feature vector at spatial position t in the reshaped feature map, consisting of C channels. In parallel, the backward GRU (GRU_{bwd}), shown in dark pink, moves in reverse—from bottom-right to top-left—starting from its initial hidden state h_{N+1}. It is defined as:

$$\overleftarrow{h}_t = \text{GRU}_{\text{bwd}}(\mathbf{P}_t, \overleftarrow{h}_{t+1}), \quad t = N, \ldots, 1 \tag{4.3}$$

It produces backward hidden states $\overleftarrow{h}_N, \overleftarrow{h}_{N-1}, \ldots, \overleftarrow{h}_1$, where each hidden state $\overleftarrow{h}_t$ depends on the next hidden state $\overleftarrow{h}_{t+1}$ and the current input $\mathbf{P}_t$.

At each spatial position $t = 1, \ldots, N$ (where $N = H' \times W'$), the forward and backward hidden states are concatenated to form the final hidden state:

$$\mathbf{h}_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \in \mathbb{R}^{2C} \tag{4.4}$$

These concatenated hidden states, ($\mathbf{h}_t$), highlighted in blue between the forward and backward GRU blocks in Fig. 4.4, provide a comprehensive bidirectional representation of spatial dependencies across the feature map.

To address the increased dimensionality introduced by concatenation (doubling the channel size from C to $2C$), a 1×1 convolutional layer is applied to project the combined features back to dimension C. As previously defined $\mathbf{F}_{\text{GRU}_i} \in \mathbb{R}^{H'W' \times 2C}$ is the output of the Bi-GRU module (see Eq. 4.1). The dimensionality is reduced by applying:

$$\mathbf{F}'_{\text{GRU}_i} = \text{Conv}_{1\times1}(\mathbf{F}_{\text{GRU}_i}) \in \mathbb{R}^{H'W' \times C} \tag{4.5}$$

The resulting Bi-GRU-enhanced features are then combined with the original ResNet features using a learnable weighted fusion mechanism, defined as:

$$\mathbf{F}_{\text{Combined}} = \alpha_i \cdot \mathbf{F}'_{\text{GRU}_i} + (1 - \alpha_i) \cdot \mathbf{F}_{\text{ResNet}_i} \tag{4.6}$$

where α_i is a trainable parameter associated with the i-th ResNet layer, and as defined in Eqs. 4.5 and 4.1, respectfully, $\mathbf{F}'_{\text{GRU}_i}$ represents the projected output of the Bi-GRU module at the i-th Bi-GRU module, and $\mathbf{F}_{\text{ResNet}_i}$ corresponds to the original feature vector from the i-th ResNet layer. This dynamic balancing mechanism adaptively integrates local

feature extraction from ResNet with the global contextual representations provided by Bi-GRUs, enhancing spatial feature modeling for downstream tasks.

4.3.3 Bottleneck Feature Refinement

The bottleneck layer serves as a pivotal transition between the encoder and decoder in the proposed architecture. It processes the output of the deepest ResNet layer, enriched by GRU-enhanced features. To further refine these features, ASPP [31], which is followed by CBAM [173], is strategically integrated into the bottleneck.

In histopathology, where tumors exhibit significant variability in size and shape, multi-scale feature extraction is essential for improving detection and localization accuracy. Conventional convolutional networks often struggle to generalize across different tumor sizes. Small metastases may be overlooked, while larger tumor regions may dominate the feature space. ASPP [31] mitigates these challenges by introducing multiple receptive fields, allowing the model to detect structures at varying scales. This ability enhances the robustness of deep learning models against tumor size variations and helps preserve spatial coherence, which is critical in metastasis detection.

As shown in Fig. 4.2, the proposed model integrates ASPP [31] at the bottleneck layer, a decision motivated by the need to refine high-level feature representations before upsampling in the decoder. The bottleneck layer is typically considered the stage where the deepest feature maps from the encoder, enriched by GRU-enhanced spatial dependencies, encode high-level semantic representations with broader contextual understanding. Placing ASPP [31] at this stage ensures that multi-scale feature extraction occurs when the model has already captured complex spatial relationships, making it particularly effective for improving localization accuracy.

In addition, refining features at the bottleneck enhances the encoder's ability to aggregate high-level contextual information with a large receptive field, which is critical for accurate spatial reconstruction during decoding [132, 119].

In the newly proposed weakly supervised learning approach, the model relies on

metastatic classification and tumor ratio regression rather than pixel-wise annotations.

Following ASPP, the feature maps undergo further refinement through the Convolutional Block Attention Module (CBAM) [173]. While ASPP extracts multi-scale features, CBAM enhances spatial and channel attention, addressing the insufficient focus of conventional CNNs on tumor-relevant regions, prioritizing diagnostically relevant regions and suppressing irrelevant background information. The following subsections provide a detailed explanation of the bottleneck layer components, including CBAM [173] and the Dual-Task Classification and Regression module.

4.3.4 CBAM in WSI Processing

WSIs present unique challenges due to their large scale, high resolution, and complex tissue variations. Accurate metastasis detection within these images requires both contextual awareness and fine-grained spatial discrimination. In the proposed model, to capture the most informative channel and spatial features of tumors, an advanced attention mechanism that sequentially applies *channel* and *spatial attention* to enhance feature representation, *i.e.*, CBAM [173], is adopted. Unlike conventional convolutional layers that process features uniformly, CBAM adaptively refines feature maps by emphasizing the most informative channels and spatial locations. This mechanism improves the network's ability to focus on diagnostically relevant regions while suppressing irrelevant or noisy information, making it particularly well-suited for WSI analysis in metastasis detection.

As shown in Fig. 4.5, CBAM [173] consists of two key components: **channel attention** and **spatial attention**. The channel attention module selectively amplifies the most important feature channels by computing global descriptors through average and max pooling. These descriptors are processed through a multi-layer perceptron (MLP) to generate attention weights, which modulate the feature channels accordingly. This selective refinement enhances the model's ability to focus on salient features, which can aid in distinguishing between different tissue patterns, including those associated with metastasis.

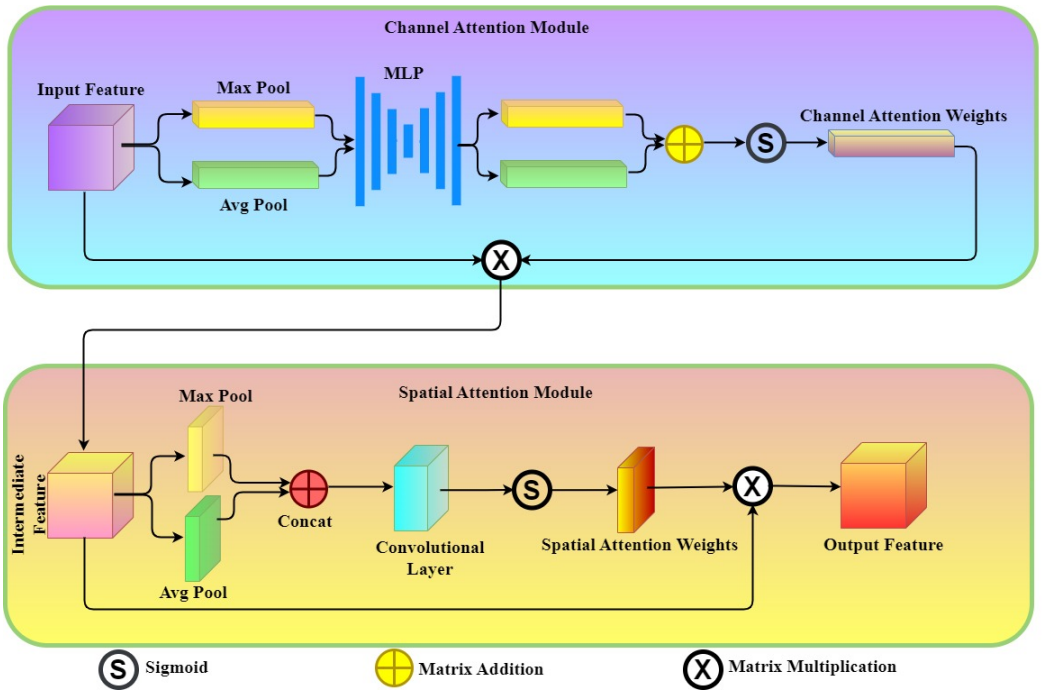


Figure 4.5 : Illustration of the CBAM module integrated into the proposed framework. The top section shows the **Channel Attention Module**, which refines the input feature map by emphasizing informative feature channels. The bottom section depicts the **Spatial Attention Module**, which further enhances the intermediate features by focusing on important spatial regions. The input feature map is progressively refined to produce the final output feature map, improving metastasis localization.

Following this refinement, the spatial attention module identifies key regions within the spatial dimensions of the feature map. Aggregating information from different channels through average and max pooling captures both global and local spatial patterns. The pooled information is concatenated and processed by a convolutional layer, generating spatial attention weights that emphasize regions with distinct patterns, such as variations in cell density, tissue texture, and structural irregularities. These regions are assigned higher weights because they exhibit features that deviate from normal tissue characteristics, making them more relevant for detecting metastatic growth. By selectively enhancing these

diagnostically relevant regions, spatial attention improves the model's ability to localize tumor areas, supporting more accurate differentiation between normal and cancerous tissue in histopathology images [173, 47].

As shown in Fig. 4.2, in the proposed model, CBAM [173] is strategically placed in the bottleneck layer, where feature maps contain rich semantic information but low spatial resolution, to process the multi-scale features generated by ASPP [31]. This step integrates the global contextual information from ASPP [31] with CBAM's attention-driven refinement, resulting in features enriched with both spatial and channel relevance. As illustrated in Fig. 4.5 CBAM sequentially applies channel and spatial attention to enhance these features. Let $\mathbf{F}_{\text{input}} \in \mathbb{R}^{2048\times H/32\times W/32}$ denote the Input Feature, which now corresponds to the output of the ASPP module [31]. First, channel attention is applied:

$$\mathbf{F}_{\text{inter}} = \mathcal{M}_{\text{c}}(\mathbf{F}_{\text{input}}) \otimes \mathbf{F}_{\text{input}} \tag{4.7}$$

where $\mathcal{M}_{\text{c}}(\cdot)$ denotes the Channel Attention Weights and $\otimes$ indicates element-wise multiplication. The result is the Intermediate Feature, $\mathbf{F}_{\text{inter}}$, which is then refined by spatial attention:

$$\mathbf{F}_{\text{output}} = \mathcal{M}_{\text{s}}(\mathbf{F}_{\text{inter}}) \otimes \mathbf{F}_{\text{inter}} \tag{4.8}$$

where $\mathcal{M}_{\text{s}}(\cdot)$ is the Spatial Attention Weights. The channel attention map, $\mathcal{M}_{\text{c}}$, is generated by modeling inter-channel relationships using global pooling operations followed by a shared transformation. The spatial attention map, $\mathcal{M}_{\text{s}}$, is computed by capturing spatial dependencies from the intermediate feature using aggregated descriptors. For full architectural details, the reader is referred to CBAM [173]. The final Output Feature, $\mathbf{F}_{\text{output}}, \in \mathbb{R}^{2048\times H/32\times W/32}$ preserves the original dimensions and contains enhanced semantic information for subsequent reconstruction in the decoder.

Thus, by integrating CBAM [173] in the bottleneck layer, the proposed model leverages both global semantic refinement and fine-grained attention mechanisms for weakly supervised WSI analysis. The synergy between ASPP [31] and CBAM [173] supports multi-scale feature refinement and attention-guided localization while remaining compu-

tationally feasible.

4.3.5 Dual-Task Framework and Loss Computation

The proposed model incorporates a dual-task learning framework that simultaneously performs classification and regression tasks, helping train the model for WSI-level localization and detection without relying directly on dense annotations. The classification task determines whether a patch contains metastatic tissue, while the regression task estimates the tumor ratio within a patch, providing a quantitative measure of the metastatic burden. Each patch is labeled with a binary class (healthy or metastatic) and assigned a tumor ratio, which quantifies the proportion of metastatic content within the patch. The tumor ratio is computed using pixel-level annotations during dataset preparation, where the proportion of annotated tumor pixels in the patch is calculated. However, during model training, pixel-level annotations are not directly utilized. Instead, the model is trained using the patch-level labels (binary class) and the tumor ratio as the target regression value. As shown in the bottom two rows of Fig. 4.2, CLF_FC and REG_FC represent the classification and regression components, respectively. These dual tasks are performed at two levels of the architecture—within the bottleneck and at the end of the decoder—leveraging both global and high-resolution spatial features.

At the bottleneck layer, global features refined by the CBAM [173] module that follows the ASPP block, are extracted from the feature map with dimensions $(2048, H/32, W/32)$. These features, denoted by $\mathbf{F}_{\text{bottleneck}}$, undergo global average pooling (GAP), followed by fully connected layers to produce the classification and regression predictions as:

$$\hat{cl}f_{\text{bottleneck}} = \text{FC}_{\text{cls}}(\text{GAP}(\mathbf{F}_{\text{bottleneck}})), \quad \hat{tr}_{\text{bottleneck}} = \text{FC}_{\text{reg}}(\text{GAP}(\mathbf{F}_{\text{bottleneck}})). \quad (4.9)$$

Here, FCcls refers to the fully connected layer used for binary classification, and FCreg denotes the fully connected layer used for tumor ratio regression.

Meanwhile, at the stage between the fourth upsampling in the decoder and the final upsampling that produces the predicted mask, as shown in Fig. 4.2, the output high-resolution spatial features with dimensions $(64, H/2, W/2)$ are used as inputs for an ad-

ditional classification and regression subbranch. Let $\mathbf{F}_{\text{decoder}}$ denote these features. They are also passed through GAP and FC layers as:

$$\hat{clf}_{\text{decoder}} = \text{FC}_{\text{cls}}(\text{GAP}(\mathbf{F}_{\text{decoder}})), \quad \hat{tr}_{\text{decoder}} = \text{FC}_{\text{reg}}(\text{GAP}(\mathbf{F}_{\text{decoder}})). \tag{4.10}$$

This earlier stage is chosen to retain richer spatial details, as the final upsampling reduces the feature map to $(1, H, W)$, limiting the capacity for effective classification and regression. By leveraging a higher-resolution representation, the subbranch enhances prediction granularity and ensures more precise feature refinement. The outputs from the bottleneck and decoder layers are combined, using an average fusion strategy, to leverage the complementary strengths of both global and fine-grained features. The classification predictions from the bottleneck and decoder layers are denoted as $\hat{clf}_{\text{bottleneck}}$ and $\hat{clf}_{\text{decoder}}$, respectively. Similarly, the regression outputs from the bottleneck and decoder layers are represented as $\hat{tr}_{\text{bottleneck}}$ and $\hat{tr}_{\text{decoder}}$. The final classification prediction, denoted by $\hat{clf}_{\text{final}}$, is computed as follows:

$$\hat{clf}_{\text{final}} = \frac{\hat{clf}_{\text{bottleneck}} + \hat{clf}_{\text{decoder}}}{2}, \tag{4.11}$$

and the final regression, denoted by $\hat{tr}_{\text{final}}$ is computed as follows. This value represents the estimated proportion of the patch occupied by tumor cells.

$$\hat{tr}_{\text{final}} = \frac{\hat{tr}_{\text{bottleneck}} + \hat{tr}_{\text{decoder}}}{2}. \tag{4.12}$$

To optimize this dual-task framework, two distinct loss functions are employed, respectively. The classification task uses the binary cross-entropy (BCE) loss, which compares the predicted metastasis probability $\hat{clf}_i$ against the ground truth label clf_i. The BCE loss is defined as:

$$L_{\text{class}} = -\frac{1}{B}\sum_{i=1}^{B}\left[clf_i \log(\hat{clf}_{\text{i}}) + (1 - clf_i)\log(1 - \hat{clf}_{\text{i}})\right], \tag{4.13}$$

where B represents the batch size.

For the regression task, the MSE loss is employed to compare the predicted tumor ratio $\hat{tr}_i$ with the ground truth ratio tr_i. As mentioned in Section 4.3.1, the tumor ratio is used as an auxiliary regression target that guides the model toward learning the

approximate spatial extent of tumor regions without relying on pixel-wise supervision. This regression objective acts as a spatially structured supervision signal, encouraging the model to capture the extent and distribution of metastatic regions, while serving as an implicit regularization mechanism that enables boundary-aware localization without requiring explicit pixel-level annotations. The MSE is defined as follows:

$$L_{\text{reg}} = \frac{1}{B}\sum_{i=1}^{B}(tr_i - \hat{tr}_i)^2. \tag{4.14}$$

where B represents the batch size as well.

Finally, a weighted combination of these two loss functions constitutes the final objective function, allowing the model to jointly optimize both objectives:

$$L_{\text{total}} = \lambda_{\text{class}} L_{\text{class}} + \lambda_{\text{reg}} L_{\text{reg}}, \tag{4.15}$$

where λ_{class} and λ_{reg} are hyperparameters that control the relative importance of the classification and regression tasks. In the proposed approach, both tasks were assigned equal weighting ($\lambda_{\text{class}} = 0.5$ and $\lambda_{\text{reg}} = 0.5$) to ensure a balanced contribution from classification and regression objectives. Empirical evaluation on the validation set supported the use of this configuration for the reported experiments. The effect of varying these hyperparameters is further explored in the ablation study (Section 4.6).

This dual-task framework eliminates the need for pixel-level annotations during training by relying solely on patch-level labels and tumor ratios. Such an approach aligns with the weakly supervised paradigm while supporting both localization and metastasis detection. By integrating classification and regression tasks within a unified architecture, the model balances prediction performance with computational feasibility in weakly supervised metastasis analysis.

4.3.6 Decoder Architecture for High-Resolution Metastasis Localization

This section introduces the decoder component responsible for generating high-resolution localization maps used to detect and localize metastatic regions in WSI. In encoder-decoder architectures, the decoder progressively upsamples and refines low-resolution feature representations to provide high-resolution outputs for localization tasks. In this model, the

decoder reconstructs detailed metastasis localization maps by integrating transposed convolutions, skip connections, and attention mechanisms, ensuring accurate spatial detail recovery from whole slide image patches. Transposed convolutions are applied at each upsampling stage, skip connections transfer multi-scale contextual features from the encoder, and CBAM modules refine the skip connection features before fusion in the decoder.

The final output of the decoder is a pixel-wise localization map that captures spatial patterns associated with tumor regions. Although no pixel-level segmentation loss is applied, the decoder's features contribute to auxiliary classification and regression branches, which guide learning and help the model localize metastases under weak supervision, as described in Section 4.3.5. Together, these components allow the model to produce high-resolution maps that support accurate tumor localization without requiring pixel-wise annotations. Further details on these components are discussed in the subsequent parts of this section.

As illustrated in Fig. 4.1, the decoder follows the bottleneck stage, which consists of ASPP and CBAM. As detailed in Fig. 4.2, the decoder begins immediately after the CBAM module in the bottleneck. The transposed convolutional upsampling operations, denoted as $\mathrm{TC_i}$ (for $i = 1$ to 5), are highlighted in purple. Meanwhile, in the middle of Fig. 4.2, the 1×1 convolution layers in the skip connections are shown in pink, followed by CBAM blocks in blue. These components work collectively to reconstruct high-resolution localization maps from the low-resolution features produced by the bottleneck layer. This design supports more precise tumor localization and improved boundary refinement while remaining computationally feasible.

The upsampling process progressively reconstructs the localization map by merging feature maps from the previous decoder layer with CBAM-refined skip connections [132] from the corresponding earlier ResNet + GRU layer. Let $\mathbf{U}_{i-1} \in \mathbb{R}^{2C' \times H/2 \times W/2}$ denote the input feature map from the previous decoder layer, and $\mathbf{F}_i \in \mathbb{R}^{C' \times H' \times W'}$ the corresponding encoder feature map. At each step, the decoder performs three operations: upsampling, refinement, and fusion. First, the decoder feature map from the previous stage, denoted

as $\mathbf{U}_{i-1} \in \mathbb{R}^{2C' \times H/2 \times W/2}$, is upsampled using a transposed convolution to produce an intermediate feature map of shape $\mathbf{U}_i^{\text{up}} \in \mathbb{R}^{C' \times H' \times W'}$:

$$\mathbf{U}_i^{\text{up}} = \text{TransposedConv}(\mathbf{U}_{i-1}) \tag{4.16}$$

Simultaneously, the corresponding encoder feature map $\mathbf{F}_i$ is processed through a 1×1 convolution and refined by CBAM [173] to generate the skip connection:

$$\mathbf{F}_i^{\text{ref}} = \text{CBAM}(\text{Conv}_{1\times1}(\mathbf{F}_i)) \tag{4.17}$$

The result, $\mathbf{F}_i^{\text{ref}}$, serves as the CBAM-refined skip connection at stage i. The final output of the decoder stage is obtained by fusing the upsampled decoder feature with the CBAM-refined skip connection through element-wise addition:

$$\mathbf{U}_i = \mathbf{U}_i^{\text{up}} + \mathbf{F}_i^{\text{ref}} \tag{4.18}$$

As illustrated in Fig. 4.6, each stage of the decoder follows a consistent structural pattern, combining upsampled decoder features with CBAM-refined skip connections, while maintaining coherence in spatial and channel dimensions.

In the first decoder stage (u_1), the input feature map from the bottleneck, denoted as $\mathbf{U}_0 \in \mathbb{R}^{2048 \times H/32 \times W/32}$, is upsampled using a transposed convolution to produce $\mathbf{U}_1^{\text{up}} \in \mathbb{R}^{1024 \times H/16 \times W/16}$. The corresponding encoder feature map $\mathbf{F}_3 \in \mathbb{R}^{1024 \times H/16 \times W/16}$, extracted from Layer 3 of the ResNet-GRU encoder, is refined using a 1×1 convolution followed by CBAM [173], yielding the skip connection $\mathbf{F}_3^{\text{ref}}$. The output of this stage is obtained by element-wise addition of the upsampled decoder feature and the refined skip connection:

$$\mathbf{U}_1 = \mathbf{U}_1^{\text{up}} + \mathbf{F}_3^{\text{ref}}. \tag{4.19}$$

This step initiates the reconstruction of spatial detail in the decoder, while preserving the high-level semantic information captured in the bottleneck. It establishes the foundation for integrating deeper contextual features with progressively finer resolution in subsequent stages.

In the second decoder stage (u_2), the input feature map $\mathbf{U}_1 \in \mathbb{R}^{1024\times H/16\times W/16}$ is upsampled using a transposed convolution to produce $\mathbf{U}_2^{\text{up}} \in \mathbb{R}^{512\times H/8\times W/8}$. The corresponding encoder feature $\mathbf{F}_2 \in \mathbb{R}^{512\times H/8\times W/8}$, extracted from Layer 2 of the ResNet-GRU encoder, is refined by a 1×1 convolution followed by CBAM [173] to produce the skip connection $\mathbf{F}_2^{\text{ref}}$. The final fused feature map is then computed as:

$$\mathbf{U}_2 = \mathbf{U}_2^{\text{up}} + \mathbf{F}_2^{\text{ref}}. \tag{4.20}$$

This stage increases spatial resolution while incorporating intermediate-level contextual information from the encoder.

During the third decoder stage (u_3), the input feature map $\mathbf{U}_2 \in \mathbb{R}^{512\times H/8\times W/8}$ is upsampled using a transposed convolution to produce $\mathbf{U}_3^{\text{up}} \in \mathbb{R}^{256\times H/4\times W/4}$. The corresponding encoder feature $\mathbf{F}_1 \in \mathbb{R}^{256\times H/4\times W/4}$, extracted from the ResNet-GRU block in Layer 1, is refined using a 1×1 convolution followed by CBAM [173], resulting in the skip connection $\mathbf{F}_1^{\text{ref}}$. The final output $\mathbf{U}_3$ is obtained by fusing $\mathbf{U}_3^{\text{up}}$ and $\mathbf{F}_1^{\text{ref}}$ through element-wise addition:

$$\mathbf{U}_3 = \mathbf{U}_3^{\text{up}} + \mathbf{F}_1^{\text{ref}}. \tag{4.21}$$

During the fourth decoder stage (u_4), the input feature map $\mathbf{U}_3 \in \mathbb{R}^{256\times H/4\times W/4}$ is upsampled using a transposed convolution to produce $\mathbf{U}_4^{\text{up}} \in \mathbb{R}^{64\times H/2\times W/2}$. The corresponding encoder feature $\mathbf{F}_0 \in \mathbb{R}^{64\times H/2\times W/2}$, extracted from the earliest ResNet-GRU block, is refined using a 1×1 convolution followed by CBAM [173] to generate the skip connection $\mathbf{F}_0^{\text{ref}}$. The final output $\mathbf{U}_4$ is then obtained by fusing $\mathbf{U}_4^{\text{up}}$ and $\mathbf{F}_0^{\text{ref}}$ via element-wise addition:

$$\mathbf{U}_4 = \mathbf{U}_4^{\text{up}} + \mathbf{F}_0^{\text{ref}}. \tag{4.22}$$

At this point, the decoder's resolution reaches half of the original input size, allowing the integration of shallow, low-level features that preserve fine spatial structures.

Finally, the feature map from the last upsampling step, $\mathbf{U}_4 \in \mathbb{R}^{64\times H/2\times W/2}$, is further upsampled using a transposed convolution to produce the final output $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$, where 1 represents the number of output channels required for localization. This final

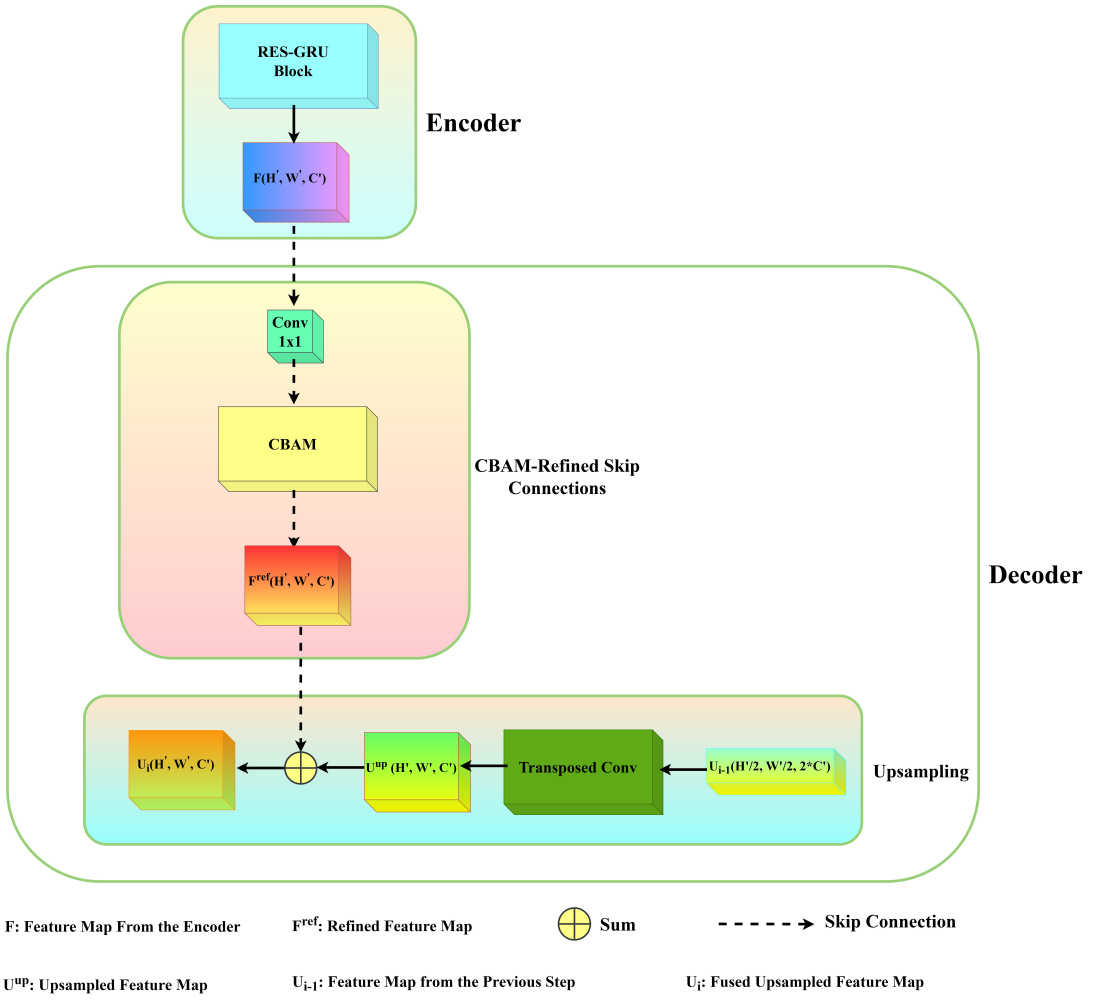


Figure 4.6 : Detailed illustration of a single decoder stage in the proposed architecture.

upsampling stage restores the spatial resolution to match the original input size while preserving the fine-grained spatial and contextual information accumulated throughout the decoder.

By aligning the output resolution with the input dimensions, this step enables high-resolution localization of metastases within the input region. The design ensures that fine structural cues—refined at each decoding stage—are retained in the final prediction, supporting precise and high-resolution output generation under weak supervision.

A detailed illustration of a stage in the decoder is shown in Fig. 4.6, demonstrating

the flow of feature refinement and fusion. The feature map from the encoder (ResNet + Bi-GRU) is passed through the skip connection and refined by the CBAM module. Simultaneously, the feature map from the previous decoder stage, with reduced spatial resolution and increased channels, is upsampled via transposed convolution. For enhanced feature representation, the upsampled decoder feature and the CBAM-refined skip connection feature are combined through element-wise addition, producing an upsampled, refined feature map that serves as the input to the next decoder stage. The final output of the decoder is a pixel-wise localization map that serves as a metastasis likelihood or density map. Decoder features are supervised using weakly labeled data through auxiliary classification and regression heads that predict patch-level metastasis labels and tumor ratios. These outputs are optimized using binary cross-entropy and mean squared error losses, respectively. During backpropagation, these task-specific losses provide learning signals that flow through the decoder, effectively shaping the spatial predictions of the network under weak supervision.

4.4 Implementation and Evaluation

The proposed method was implemented using the PyTorch framework, and experiments were conducted on a single NVIDIA Tesla V100-PCIE-32GB GPU. For optimization, the Adam optimizer was utilized with an initial learning rate of 1×10^{-4}. To support the learnable fusion between GRU and ResNet outputs, where each layer's contribution is controlled by a trainable scalar weight α_i, a separate learning rate of 1×10^{-3} was assigned to these fusion weights. This was implemented using PyTorch's `param_groups` to define a distinct learning rate for the α_i parameters within the same optimizer. A weight decay of 5×10^{-4} was used to regularize the model, and ReLU was used as the activation function in all convolutional layers throughout the encoder and decoder, except for the final output layer.

All hyperparameters were selected using validation data only, with the official CAMELYON16 test set kept strictly isolated throughout model development and evaluation. No

test-set performance was used for model selection or parameter adjustment.

This validation-based selection covered the main training hyperparameters used throughout this chapter, including the optimizer learning rates, weight decay, and the relative loss weights for classification and tumor ratio regression. Once selected, these values were kept fixed for all reported experiments in this chapter.

For the proposed method, the same post-processing configuration was applied unchanged across the patch-split, WSI-level split, and full-training-set experiments.

The primary aim of this work is to detect and localize metastases in gigapixel pathology images of breast cancer. Given the large size of WSIs, they are first divided into smaller, more manageable patches before being input into the model. Each metastatic WSI contains both healthy and unhealthy regions.

For training, 25 metastatic WSIs and 15 healthy WSIs were randomly selected from the CAMELYON16 training dataset. A total of 200,000 patches, each of size 256×256, were randomly cropped from these 40 WSIs, with equal numbers of metastatic and non-metastatic patches (100,000 each). During each training iteration, the model produces five outputs for each input patch: a 256×256 localization mask at the decoder's output, two classification outputs from the bottleneck and decoder, and two regression outputs aimed at estimating tumor ratios.

For inference, the model processes larger input images of size 2048×2048, constrained only by GPU memory. Each test WSI is divided into overlapping regions of 2048×2048 with an overlap size of 128 pixels. These regions are sequentially fed into the model to generate corresponding 2048×2048 localization masks, which are then stitched together to produce the final mask for the entire WSI. The resulting WSI mask is a probability map where each pixel is assigned a score representing the likelihood of metastasis. The final whole-slide probability map is reconstructed by stitching together the overlapping tile-level predictions and averaging the overlapping areas to ensure smooth transitions across tile boundaries.

This probability map represents a dense map, where each pixel is assigned a metastasis likelihood score between 0 and 1. It is important to note that these probability scores are derived from the model's soft outputs and may not represent well-calibrated posterior probabilities. Deep neural networks are known to produce overconfident predictions, particularly in ambiguous or out-of-distribution regions. Therefore, the probability map should be interpreted as a relative confidence measure rather than a fully reliable estimate of predictive certainty. To generate a binary localization mask, a fixed probability threshold of 0.5 is applied to the probability map. This converts the dense probability map into a binary localization mask suitable for deterministic morphological refinement.

Morphological operations are then applied deterministically to refine the predicted masks. Specifically, morphological opening with a 3×3 square structuring element is used to suppress isolated noise artifacts. This is followed by morphological closing with a 5×5 square structuring element to bridge narrow discontinuities introduced by patch-wise prediction and stitching. Hole filling is subsequently applied to improve the topological consistency of predicted metastatic regions.

Finally, connected component analysis is employed to eliminate spurious detections. In this step, connected components with an area smaller than 100 pixels at $10\times$ magnification are discarded, as they are unlikely to correspond to diagnostically meaningful metastases. These post-processing steps suppress small isolated artifacts and regularize the predicted masks while preserving the overall lesion geometry.

All post-processing parameters are specified explicitly and applied deterministically. The probability threshold, structuring element sizes, and minimum connected-component area were selected using validation data only and then kept fixed across all experiments in this chapter. No post-processing parameters were tuned, adjusted, or selected based on performance on the official CAMELYON16 test set, which remained strictly isolated during evaluation.

For FROC evaluation, the probability map is first thresholded and morphologically refined to produce a binary mask indicating predicted tumor regions. Connected compo-

nents in this mask are treated as candidate lesion localizations. Each candidate region is then assigned a confidence score equal to the maximum probability value within that region, taken from the corresponding raw probability map. These confidence scores are used to rank lesion candidates when computing the FROC curve.

Because lesion-level FROC scoring based on the maximum pixel probability may be sensitive to isolated high-confidence outliers, a small sensitivity analysis was also conducted. In addition to the primary max-probability scoring rule, each predicted connected region was alternatively scored using the average of its top-K pixel probabilities, with $K = 50$. The resulting FROC performance remained comparable to that obtained with max-based scoring, indicating that the reported localization performance is not driven by isolated single-pixel responses and that the lesion-ranking procedure is reasonably stable.

Dice [38] and IoU [71] are computed between the thresholded WSI probability maps and the ground-truth tumor masks to quantify boundary-level spatial agreement (Table 4.5).

For the AUC metric, prior CAMELYON16 approaches extracted handcrafted features from the predicted heatmaps and used additional classifiers to predict the probability of a WSI containing tumors. In contrast, similar to the approach described in Chapter 3, the proposed method uses a simpler slide-level scoring strategy. A single WSI-level metastasis score is obtained by averaging the predicted probabilities over the detected foreground regions in the reconstructed probability map. This score is then used to compute the slide-level AUC. By avoiding additional feature extraction or learning-based post-processing, the method remains computationally efficient while still producing competitive AUC performance.

Together, these evaluation procedures provide a consistent assessment of both lesion-level localization and slide-level classification performance under the same inference and post-processing pipeline.

4.4.1 Dataset

The model was trained and evaluated using the CAMELYON16 challenge dataset [20]. For training, a subset of 40 WSIs (25 metastatic and 15 healthy) was selected from the original 270 training images. Testing, consistent with the previous approach, was conducted on all 130 WSIs in the test set. This enables comparison with previously reported CAMELYON16 results.

4.4.2 Evaluation Metrics

To benchmark the results against previously reported methods for metastasis detection, and similar to the approach used in Chapter 3, the CAMELYON16 competition metrics were employed. FROC is used to assess lesion-level localization performance, while AUC evaluates the model's effectiveness in whole-slide classification.

In addition to the competition metrics, overlap-based metrics are reported to quantify spatial agreement between predicted tumor regions and ground-truth annotations. Specifically, the Dice coefficient [38] and Intersection over Union (IoU) [71] are computed from the thresholded WSI probability maps. Detailed boundary-level results are presented in Section 4.5.5.

4.5 Experiment Results

4.5.1 Visual Evaluation

In this section, qualitative results are presented to illustrate the performance of the proposed method. Fig. 4.7 serves as a visual demonstration of the results obtained from the proposed method, alongside their comparison with the ground truth annotations for WSIs.

In Fig. 4.7, the first column presents the original WSIs, where tumor regions are outlined with polygons. The second column focuses on the ground truth mask, showcasing the tumor regions as delineated by expert pathologists. Finally, the third column reveals

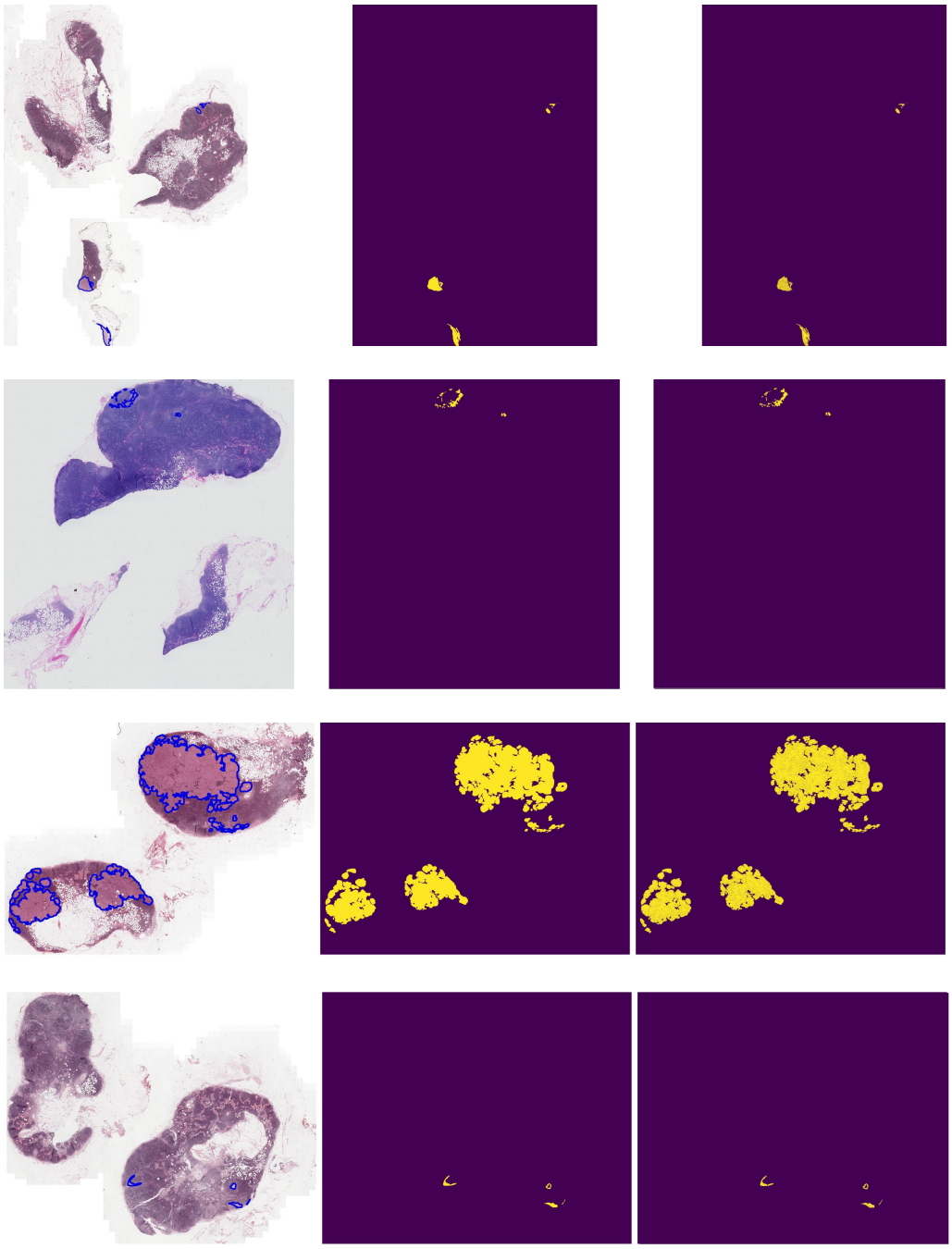


Figure 4.7 : Comparison between ground truth annotations and the predicted binary metastasis masks generated by the proposed method.

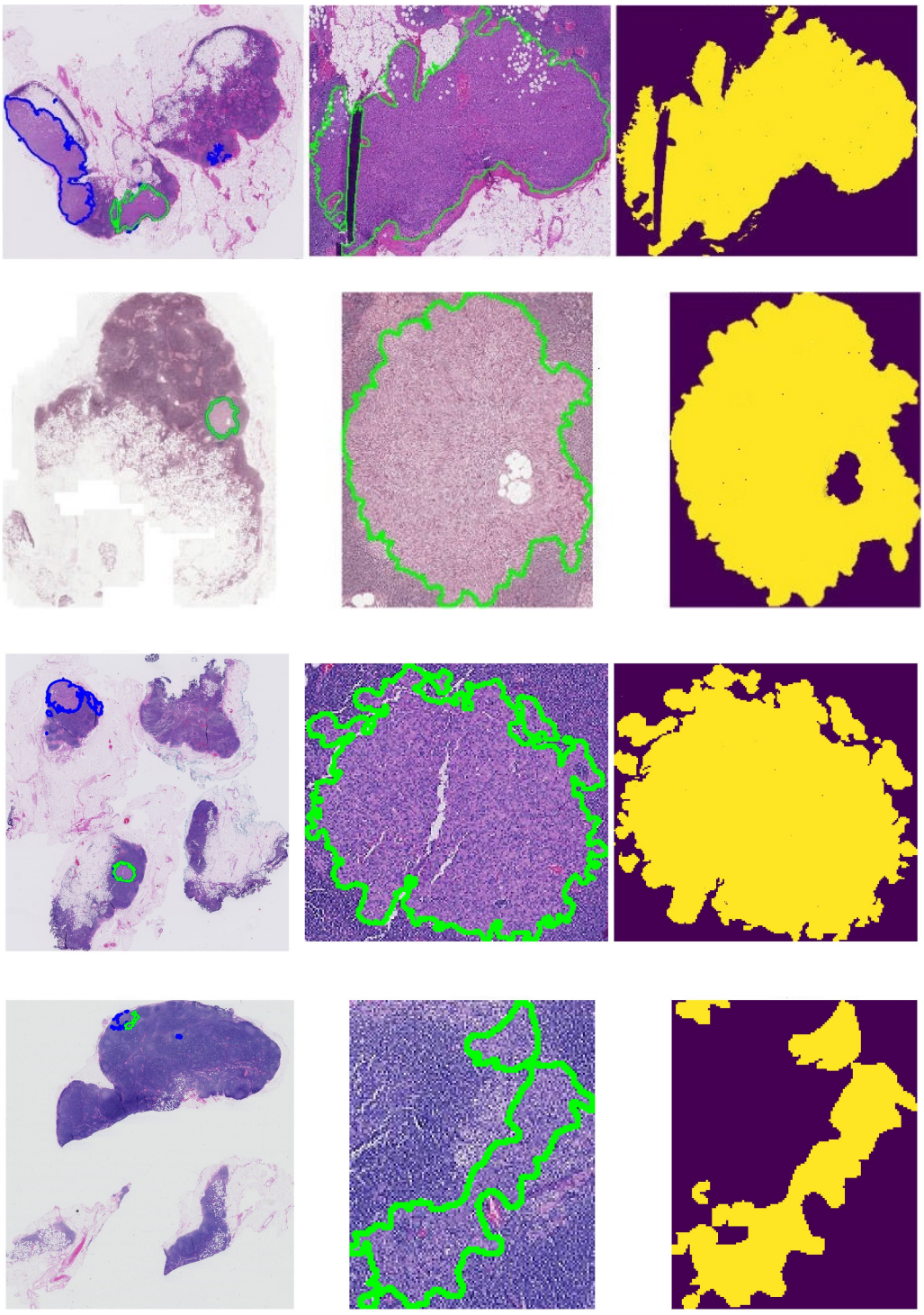
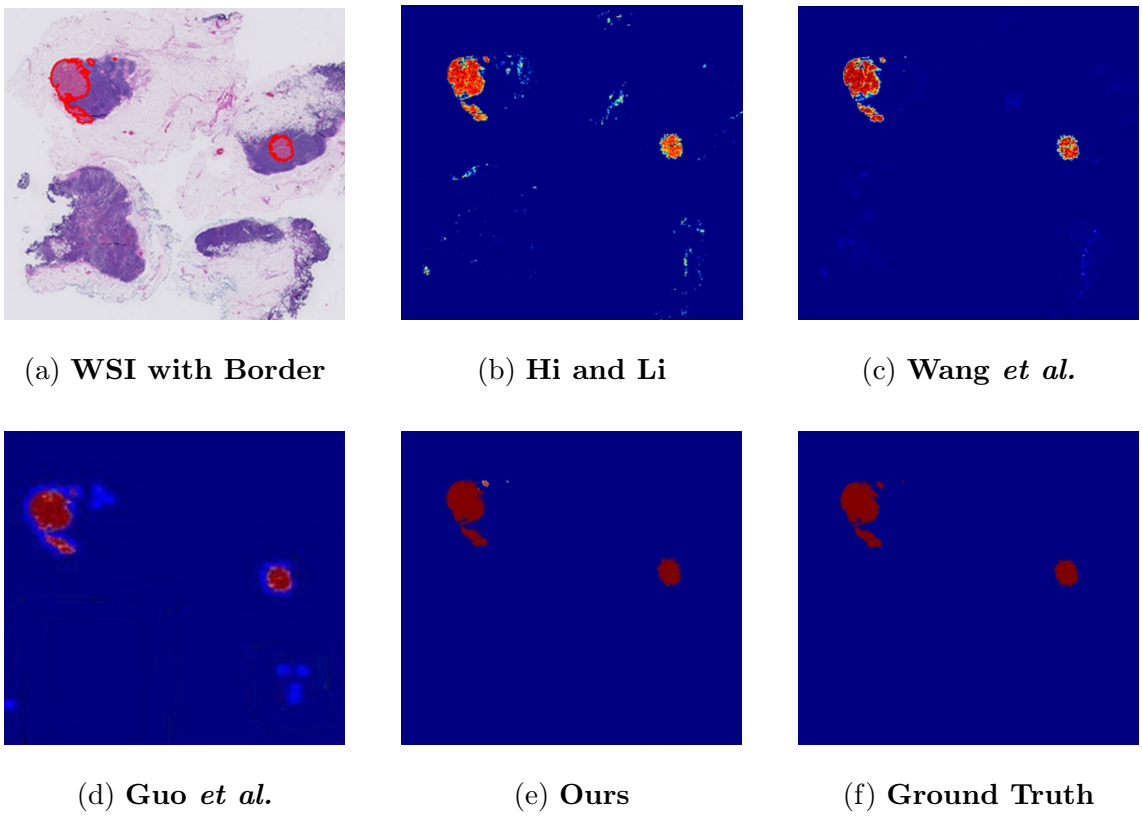


Figure 4.8 : Detailed analysis of a selected region of interest (ROI) within a WSI. The first column shows the input WSI, with the selected region highlighted by a green border. The second column presents the corresponding ground truth mask at higher resolution, while the third column shows the binary mask predicted by the proposed method.

the mask outputs generated by the proposed method, offering a direct comparison with the ground truth to validate the model's performance.



(a) **WSI with Border** (b) **Hi and Li** (c) **Wang *et al.***

(d) **Guo *et al.*** (e) **Ours** (f) **Ground Truth**

Figure 4.9 : A comparison of tumor localization results across different methods. (a) WSI with metastasis regions outlined. (b) Hi and Li [63], (c) Wang *et al.* [169], and (d) Guo *et al.* [57] show probability maps, where warmer colors indicate higher tumor likelihood. (e) Probability map generated by the proposed method, visualized using a red–blue color scheme for consistent comparison, and (f) ground truth annotation.

To further underscore the model's capabilities, Fig. 4.8 provides a closer examination of a specific region of interest. The green border in the first column highlights the selected region for zoomed-in analysis, offering a closer look at the intricate boundaries of tumor regions. The second column reveals the ground truth mask for the selected area, emphasizing the metastasis regions as defined by expert annotations. The third column showcases the mask outputs generated by the proposed model, illustrating the model's

ability to accurately delineate tumor boundaries, even in smaller or less distinct areas. This zoomed-in perspective highlights how effectively the model identifies subtle tumor margins and resolves challenges posed by less prominent borders.

For each input WSI, patches are extracted sequentially to ensure comprehensive coverage of the tissue region. The model generates a probability map for each patch, representing the likelihood of metastatic tissue. These patch-level probability maps are then merged to construct a WSI-level probability map. A binary tumor mask is obtained by thresholding the aggregated WSI probability map. This process preserves spatial consistency across patches and enables spatially coherent and boundary-aware localization of metastatic regions.

Additionally, for a more direct comparison, Fig. 4.9 illustrates the performance of various approaches for tumor localization on a WSI from the test dataset. This visualization illustrates qualitative differences between the proposed method and previously reported approaches for tumor localization.

Notably, the robustness of the model is evident despite being trained on only one-sixth of the available dataset, using only 40 WSIs for training. Even with this limited training data, the model demonstrates strong performance in detecting and delineating metastatic regions. These results highlight the capability of the proposed framework to perform metastasis detection and localization in whole-slide histopathology images under reduced training data conditions.

4.5.2 Objective Evaluation

The proposed method integrates WSI classification and localization seamlessly within a single end-to-end framework, eliminating the need for an aditional classification model. Unlike other approaches, such as Wang *et al.* (2016), which train dedicated networks for the classification of features extracted from predicted masks, the method significantly reduces computational costs and training time.

To mitigate inter-centre staining variability and improve robustness to domain shift,

Macenko stain normalization [107] was applied to extracted 256×256 tiles at $10\times$ magnification prior to model input, using a fixed reference template selected from the CAMELYON16 training set and kept constant across all experiments.

For each WSI, localization masks are constructed by merging the predicted masks of the ROIs.

Table 4.1 : Comparison with representative prior methods on CAMELYON16, including the proposed model at $10\times$ using 40 WSIs.(official test set)

Authors	**Architecture**	**AUC (%)**	**FROC (%)**	**Resolution**	**WSIs**
Pathologist	N/A	96.6	72.4	40X	N/A
Wang *et al.* (2016) [169]	GoogLeNet	**99.4**	80.7	40X	270 WSIs
CULabc	VGG-16	94.0	70.3	40X	27 WSIs
Liu *et al.* (2017) [102]	ResNet	97.6	76.0	40X	270 WSIs
Guo *et al.* (2019) [57]	Inception-v3 and DCNN	96.6	83.5	40X	270 WSIs
Khaliliboroujeni *et al.* (2022) [82]	ResNet-50 + Spatial & Channel Attention	97.03	84.1	10X	80 WSIs
Tourniaire *et al.* (2023) [157]	Mixedly Supervised-CLAM	98.2	-	40X	270 WSIs
He and Li (2024) [63].	Few-Shot + Reverse Distillation	90.36	50.0	40X	10 WSIs
Rui *et al.* (2024) [133]	Weakly Supervised Learning	88.42	78.63	5X	270 WSIs
Li *et al.* (2024) [95]	DeeplabV3+	97.4	-	10X	270 WSIs
Ours (WSI-level split, k-fold)	ResNet-50 + Bidirectional GRUs [33]+ CBAM_ASPP	99.0 ± 0.25	86.2 ± 0.5	10X	40 WSIs

For the WSI classification task, the proposed method builds upon the techniques introduced in Chapter 3, distinguishing it from other previous approaches. It does not require training an additional model from scratch using features extracted from the predicted mask. This approach reduces redundant efforts and minimizes time and resource consumption by incorporating classification within the main framework. After localization, each WSI mask is graded by averaging the probabilities of foreground pixels across all detected regions. This approach achieves a mean AUC of 99.0 ± 0.25 under WSI-level k-fold cross-validation (Table 4.1), indicating strong classification performance under a stain-normalized, leakage-resistant evaluation protocol. Notably, the proposed weakly supervised model accomplishes this using less than one-sixth of the CAMELYON16 training data, highlighting its efficiency.

In the tumor region localization task, the method achieves a mean FROC of 86.2 ± 0.5 under WSI-level k-fold cross-validation (Table 4.1).

To minimize data acquisition and processing costs, recent studies have relied on less annotated data or lower magnifications in WSIs. Despite implementing both, the proposed method achieves competitive accuracy and localization performance under reduced-data and lower-magnification settings.

These outcomes, summarized in Table 4.1, underscore the effectiveness and efficiency of the method for both detection and localization tasks. Results are reported as stated in the original publications; experimental settings (e.g., magnification level and number of WSIs) vary across studies and are therefore not directly comparable. Notwithstanding these differences, the proposed method achieves strong AUC/FROC at $10\times$ using only 40 WSIs under a stain-normalized WSI-level protocol.

4.5.3 Sensitivity Analysis of Lesion Scoring Strategies

In the proposed framework, lesion-level confidence scores for FROC evaluation are derived from the maximum tumor probability within each detected connected region. This strategy allows each candidate lesion to be ranked according to the most confident tumor response predicted within that region. However, relying solely on the maximum pixel probability may potentially be sensitive to isolated high-probability pixels produced by the network.

To assess the sensitivity of the lesion scoring mechanism, an additional analysis was conducted using an alternative aggregation strategy. Instead of assigning the lesion score based on the maximum probability value, each detected connected region was alternatively scored using the average probability of its top-K highest scoring pixels. In this analysis, $K = 50$ pixels were used as a representative small subset of the highest-confidence responses within each candidate region. This choice provides an aggregation over multiple high-probability pixels while reducing sensitivity to isolated prediction outliers.

The use of multiple high-confidence responses for decision aggregation is conceptually

related to instance selection strategies employed in weakly supervised multiple-instance learning frameworks for whole-slide image analysis, where a subset of the most informative instances contributes to the final prediction [105]. Although the present experiment focuses on evaluation rather than training, the same principle motivates the use of a small group of high-confidence pixels when computing lesion scores.

The resulting FROC performance obtained with the alternative scoring strategy was comparable to that produced by the primary max-probability approach. This observation suggests that the overall lesion ranking is not strongly affected by the choice of the aggregation rule.

Table 4.2 : Sensitivity analysis of lesion-level scoring strategies for FROC evaluation on the CAMELYON16 dataset.

Scoring Strategy	**FROC (%)**
Maximum pixel probability	86.4
Top-50 average probability	86.1

As shown in Table 4.2, the alternative top-K averaging strategy produces a similar FROC score compared to the maximum-probability approach. This observation suggests that the reported localization performance is not highly sensitive to the specific lesion scoring rule used in the evaluation.

4.5.4 Controlled Evaluation and Robustness Analysis

To ensure a rigorous and unbiased evaluation, all primary results are reported under a stain-normalized, WSI-level validation protocol, which is designed to mitigate data leakage, selection bias, and domain shift. Specifically, dataset partitions are defined at the WSI level prior to patch extraction, and Macenko stain normalization [107] is applied tile-wise (256×256 at $10\times$) prior to model input using a fixed reference template from the CAMELYON16 training set. Unless stated otherwise, AUC and FROC are reported on

the fixed official CAMELYON16 test set using WSI-level k-fold cross-validation (here, $k = 5$), with results summarized as mean $\pm$ standard deviation across folds. For controlled comparison, all methods are evaluated at $10\times$ magnification using the same subset size (40 WSIs).

For transparency and continuity with earlier exploratory analyses, the patch-split setting is additionally reported as a single-run *reference*; however, it is not used to support the primary conclusions of this chapter.

In this context, a Vision Transformer–based multiple instance learning model (ViT-MIL) is included as a foundation-model baseline. Vision Transformers represent a class of modern backbone architectures that employ self-attention to capture long-range dependencies across instances. Transformer-based MIL models have been widely adopted in whole-slide image analysis and have demonstrated competitive slide-level classification performance on CAMELYON-style benchmarks using weak, slide-level supervision [141, 91, 46].

In this study, the ViT-MIL baseline is evaluated under the same controlled conditions as the proposed method, including $10\times$ resolution, stain normalization, and WSI-level data splitting. Evaluation focuses on slide-level AUC and lesion-level FROC metrics and does not rely on pixel-level annotations. Under this setting, the ViT-MIL baseline achieves lower classification and localization performance than the proposed framework. This difference reflects variations in model design and training objectives, as the proposed method is explicitly developed for weakly supervised metastasis detection and localization, whereas ViT-MIL primarily optimizes slide-level discrimination [141, 91].

In addition, a U-Net baseline [132] is evaluated under the same weak supervision setting as the proposed method, using only patch-level labels and tumor proportion targets. During training, no pixel-level annotations or spatial ground-truth masks are used, ensuring an identical supervision regime and a fair controlled comparison.

The results presented in Table 4.3 show consistent performance trends across the evaluated methods under the controlled settings. The proposed framework achieves higher AUC and FROC values compared to the baseline approaches, while all methods are evaluated

Table 4.3 : Controlled comparison on CAMELYON16 (official test set) at 10× using 40 WSIs with stain normalization. Primary results use WSI-level k-fold cross-validation (mean ± std across folds); the patch-split setting is reported as a single-run reference.

Method	**AUC (%)**	**FROC (%)**	**Resolution**	**WSIs**
Ours (patch-split, reference)	98.79	85.7	10×	40
Ours (WSI-level split, k-fold)	99.0 ± 0.25	86.2 ± 0.5	10×	40
ViT-MIL [141] (foundation baseline)	95.5	68.0	10×	40
U-Net (weakly supervised, baseline)	96.9	81.5	10×	40

under identical resolution, dataset size, and validation conditions. This improvement reflects the task-specific design of the proposed framework for weakly supervised metastasis detection and localization.

The controlled results demonstrate that the proposed framework maintains strong whole-slide image (WSI) classification and lesion-level localization performance under a WSI-level validation protocol designed to prevent data leakage. Under identical resolution and dataset-size constraints, the ViT-MIL baseline exhibits lower classification and localization performance. A similar performance trend is observed for the classical encoder–decoder baseline, highlighting the advantage of the proposed task-specific modeling strategy under weak supervision.

To address potential selection bias arising from training on a reduced subset of WSIs, the proposed framework is further evaluated using the full CAMELYON16 training set (270 WSIs), while keeping the architecture and evaluation pipeline unchanged. This experiment serves as a robustness check under increased training coverage, and the corresponding results are reported in Table 4.4. The improved performance observed under full-dataset training is attributed to increased data diversity, which enhances generaliza-

Table 4.4 : Performance of the proposed method using the full CAMELYON16 training set (270 WSIs) at 10× with stain normalization.

Setting	**AUC (%)**	**FROC (%)**	**Resolution**	**WSIs**
Ours (full training set)	99.5	87.1	10×	270

tion and reduces sensitivity to sampling bias.

Overall, these controlled experiments demonstrate that the proposed framework maintains strong classification and localization performance under a stain-normalized, WSI-level validation protocol and improves further with expanded training coverage, supporting robustness under reduced-data and lower-magnification settings.

4.5.5 Boundary-Level Evaluation

In addition to lesion-level localization performance measured by FROC, boundary-level agreement is evaluated using the Dice coefficient [38] and Intersection-over-Union (IoU) [71]. These overlap-based metrics quantify spatial agreement between predicted regions and ground-truth annotations at the pixel level.

Both metrics are computed on a per-WSI basis from binary masks obtained by thresholding the aggregated probability maps, and are reported as average overlap scores on the CAMELYON16 test set. Consistent with the primary evaluation protocol in this chapter, the main boundary-level results are reported under WSI-level evaluation with stain normalization.

Under WSI-level evaluation at 10× using 40 WSIs, the proposed method achieves a Dice score of 0.78 and an IoU of 0.64. When trained on the full CAMELYON16 training set (270 WSIs), boundary agreement further improves to 0.81 Dice and 0.68 IoU. For transparency, the patch-split setting is also reported as a single-run reference (Dice = 0.77, IoU = 0.63); however, it is not used to support the primary conclusions.

Table 4.5 : Boundary-level overlap performance on the CAMELYON16 test set at 10× with stain normalization. WSI-level results correspond to the primary evaluation protocol of this chapter; the patch-split row is provided as a single-run reference.

Method / Setting	**Dice**	**IoU**	**Resolution**	**WSIs**
Ours (patch-split, reference)	0.77	0.63	10×	40
Ours (WSI-level split)	0.78	0.64	10×	40
Ours (full training set)	0.81	0.68	10×	270
U-Net (weakly supervised, baseline)	0.70	0.55	10×	40

For comparison, a weakly supervised U-Net baseline, trained under identical supervision and data-splitting conditions, achieves Dice = 0.70 and IoU = 0.55. These trends are consistent with lesion-level localization performance. However, overlap-based metrics quantify spatial agreement only for detected regions and do not directly capture missed lesions.

Although the proposed framework produces probability maps for metastasis likelihood, explicit uncertainty quantification (UQ) is not incorporated in the current study. This represents a limitation for high-stakes decision support, as model confidence may not reliably reflect prediction correctness. Future work may investigate established UQ techniques, such as Monte Carlo dropout and deep ensembles, to provide calibrated confidence estimates and improve the interpretability and reliability of model outputs in clinical decision-making scenarios.

4.6 Ablation Studies

To assess the contribution of individual architectural components, ablation studies are conducted in which key modules are removed while the remaining training and inference pipeline is kept unchanged. The ablations focus on the Bi-GRUs [33] in the encoder, the

ASPP [31] module in the bottleneck, and the CBAM [173] with skip connections [132] in the decoder/upsampling stage.

All ablation experiments are conducted under the same WSI-level validation protocol and Macenko stain normalization described in Section 4.5.4, ensuring consistency with the main experimental setup. The dataset follows the specification in Section 4.4.1, and all other experimental settings are kept fixed across ablation variants to enable a controlled comparison. Ablations are compared using AUC/FROC and Dice/IoU under the same evaluation pipeline. The quantitative results are summarized in Table 4.6.

Table 4.6 : Ablation results obtained by removing individual modules while keeping the training and evaluation protocol fixed under the WSI-level setting with stain normalization.

GRUs	ASPP	CBAM [173]+ skip connections [132]	WSI Classification (AUC, %)	Region Localization (FROC, %)	Dice	IoU
✗	✓	✓	98.36	85.67	0.760	0.615
✓	✗	✓	98.05	85.38	0.752	0.606
✓	✓	✗	97.88	85.12	0.740	0.592
✓	✓	✓	99.00	86.20	0.780	0.640

4.6.1 The Effectiveness of the Bi-GRUs in the Encoder

In the proposed architecture, Bi-GRUs are integrated into the encoder stage following each ResNet layer to encourage context-aware feature encoding. To evaluate the contribution of Bi-GRUs [33], an ablation study was performed by removing all Bi-GRUs [33] from the encoder. This left only the ResNet-50 backbone while keeping all other parts of the model unchanged. Without Bi-GRUs [33], the encoder relies solely on the hierarchical features generated by ResNet, potentially losing some spatial and contextual insights.

As shown in Table 4.6, under the same WSI-level split and stain-normalized evaluation protocol, removing Bi-GRUs leads to lower AUC and FROC compared to the full configuration. A corresponding reduction is also observed in Dice and IoU, indicating a consistent performance decrease under identical training conditions.

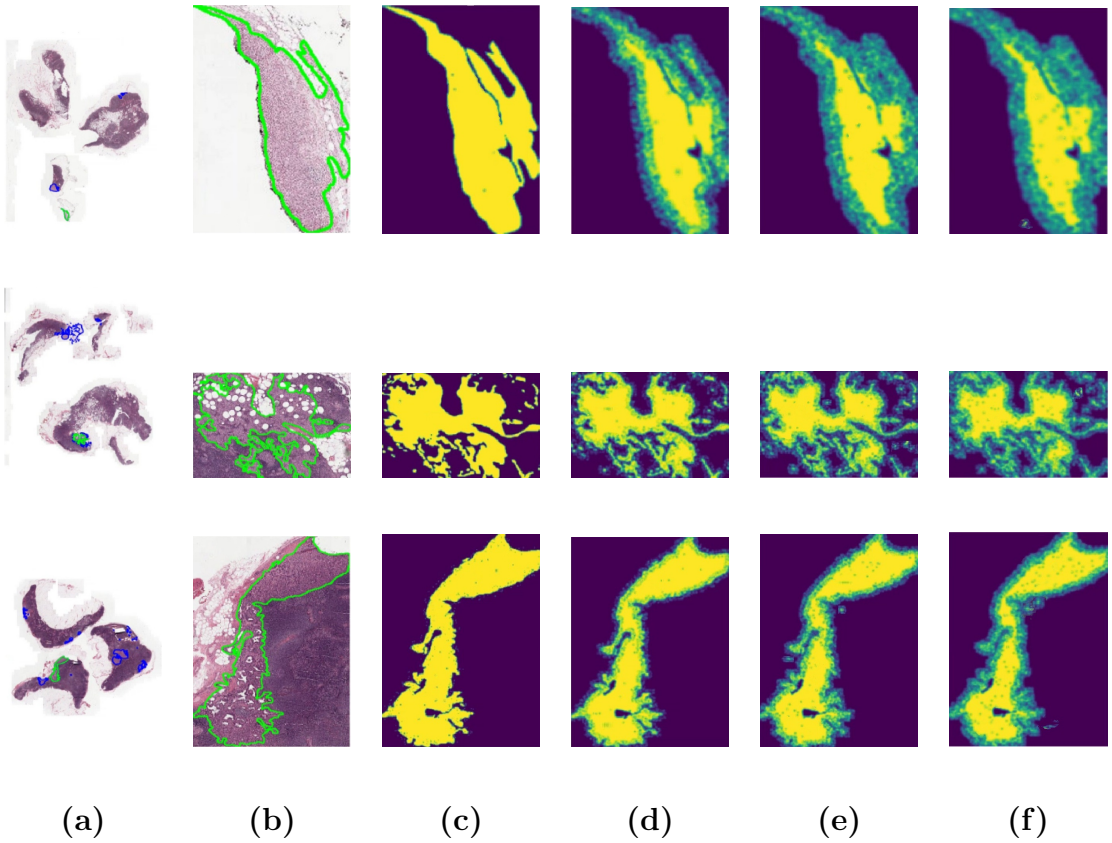


Figure 4.10 : Figure 4.10: Comparison of localization results under different model configurations. (a) Original WSI with ground truth metastasis boundaries (blue) and selected region of interest (ROI) highlighted in green. (b) Zoomed-in ROI with ground truth annotation. (c) Probability map generated by the proposed model. (d)–(f) Probability maps obtained from ablated variants of the model without Bi-GRUs, ASPP, and CBAM with skip connections, respectively, where warmer colors indicate higher tumor likelihood.

These results suggest that incorporating Bi-GRUs provides measurable performance benefits over the ResNet-only encoder within the evaluated framework.

4.6.2 The Effectiveness of ASPP in the Bottleneck

In the proposed architecture, ASPP [31] is a key component of the bottleneck layer, which employs parallel atrous convolutions with varying dilation rates to capture multi-

scale contextual information.

As shown in Table 4.6, removing ASPP under the same WSI-level split and stain-normalized protocol results in lower AUC and FROC compared to the full configuration. A corresponding decrease is observed in Dice and IoU, indicating reduced boundary-level agreement under identical training conditions.

These findings suggest that ASPP contributes positively to performance within the evaluated framework, without implying that it is uniquely necessary or optimal among all possible multi-scale designs.

4.6.3 The Effectiveness of the CBAM [173] and skip connections [132]

CBAM [173] and skip connections [132] are critical components in the proposed architecture, aimed at enhancing localization accuracy and boundary refinement. To assess their effectiveness, an ablation study is performed by removing both CBAM [173] and skip connections [132]. The dataset and experimental parameters are consistent with the original setup to allow for a fair comparison.

As reported in Table 4.6, removing CBAM and skip connections results in lower AUC and FROC compared to the full configuration. Dice and IoU also decrease relative to the complete model, indicating reduced boundary-level agreement under the same evaluation protocol.

These observations suggest that the inclusion of CBAM and skip connections contributes positively to performance within the tested configuration, while remaining limited to the empirical differences reported in Table 4.6.

In conclusion, the ablation studies indicate that each architectural component contributes to the overall performance of the proposed framework under the same training and evaluation configuration. Removing Bi-GRUs results in a consistent reduction across AUC, FROC, Dice, and IoU, suggesting that contextual feature aggregation provides measurable benefit within the encoder. Excluding ASPP leads to further degradation in both classification and lesion localization performance, consistent with reduced multi-scale con-

textual representation. The combined removal of CBAM and skip connections produces the largest performance decrease among the evaluated ablation variants, particularly in Dice and IoU, indicating reduced boundary-level agreement and spatial precision.

Overall, the quantitative results demonstrate that performance differences across variants are consistent across both lesion-level metrics (AUC/FROC) and boundary-overlap metrics (Dice/IoU). These findings support the complementary roles of contextual modeling, multi-scale aggregation, and feature refinement within the proposed architecture, without implying that this configuration is uniquely optimal among all possible alternatives.

To further illustrate these differences, Fig. 4.10 presents a visual comparison of the predicted probability maps under different ablation settings. Probability maps are used in this figure to better illustrate the impact of different architectural components on localization performance, as they reveal variations in prediction confidence that are not visible in thresholded binary masks. The first column shows the full WSI with annotation borders displayed in blue, while the zoomed Region of Interest (ROI) is highlighted in green. The second column shows the zoomed ROI with the corresponding ground-truth boundary. The third column presents the probability map generated by the full model. The subsequent columns illustrate the outputs obtained after removing Bi-GRUs, ASPP, and CBAM with skip connections, respectively.

All probability maps are visualized using an identical color scale for consistency. The full model exhibits sharper probability gradients at lesion boundaries and more compact activation regions. In contrast, the ablated variants display slightly broader transition zones and increased spatial diffusion near complex boundaries.

Collectively, the visual results are consistent with the quantitative findings reported in Table 4.6, supporting the observed differences in boundary agreement and lesion-level detection performance.

Table 4.7 : Ablation study of classification and regression loss weights ($\lambda_{\text{class}}, \lambda_{\text{reg}}$) under the WSI-level protocol with stain normalization. Results are reported using AUC (WSI-level detection), FROC (lesion localization), and boundary-level Dice/IoU. Bold indicates the best-performing configuration.

λ_{class}	λ_{reg}	AUC (%)	FROC (%)	Dice	IoU	Reason
1.0	0.0	98.41	84.96	0.744	0.596	Classification only.
0.9	0.1	98.58	85.21	0.752	0.604	Classification dominant.
0.7	0.3	98.79	85.74	0.764	0.618	Slight classification focus.
0.5	**0.5**	**99.00**	**86.20**	**0.780**	**0.640**	Balanced weighting.
0.3	0.7	98.86	85.89	0.772	0.627	Slight regression focus.
0.1	0.9	98.63	85.38	0.759	0.611	Regression dominant.
0.0	1.0	98.37	84.83	0.746	0.598	Regression only.

4.6.4 The Effectiveness of loss weightings on overall model performance

In addition, to examine the impact of loss weightings on overall model performance, an ablation study is conducted by varying the values of λ_{class} and λ_{reg} in the total loss function (Eq. 4.15). All experiments were performed under the same WSI-level split and stain normalization protocol to ensure a controlled comparison. While the classification and regression branches serve as auxiliary tasks, their contribution plays a crucial role in guiding the model toward effective WSI-level localization and detection. As shown in Table 4.7, different weighting configurations were tested, and evaluated using AUC, FROC, Dice, and IoU to assess both lesion-level and boundary-level performance.

Among the tested configurations, assigning equal weights ($\lambda_{\text{class}} = \lambda_{\text{reg}} = 0.5$) yields the strongest overall performance across the reported metrics. These findings indicate that balanced weighting between classification and tumor-rate regression objectives is beneficial under the evaluated training configuration.

4.7 Performance Efficiency Evaluation

To quantify computational cost, the runtime of the proposed framework is evaluated at both ROI and WSI levels under the same WSI-level validation and stain normalization protocol used throughout Chapter 4.

As shown in Table 4.8, the proposed method processes a single ROI of size 2816×2816 with a stride of 32 in just 0.579 minutes using a single GPU. The ROI-level runtime is within the range of previously reported CNN-based approaches, while preserving competitive localization and classification performance.

Table 4.8 : Comparison of processing time for the proposed method and representative existing methods for processing an ROI and a full WSI under comparable inference settings.

Team	**Architecture**	**ROI Runtime (min)**	**WSI Runtime (min)**
Ours	End-to-end	0.579	231.6
Pathologist	N/A	0.034	13.95
Wang *et al.* [169]	GoogLeNet	0.671	268.40
Liu *et al.* [102]	ResNet	0.729	291.60
CULab (Chinese University Lab)	VGG-16	0.979	391.60

To estimate the processing time for an average WSI, which consists of approximately 400 ROIs of size 2816×2816, the proposed model requires 231.6 minutes to complete metastasis detection.

Compared with the simpler architecture presented in Chapter 3 (2.04 minutes per WSI), the increased runtime reflects the additional architectural components introduced in Chapter 4 for enhanced contextual modeling and boundary-level refinement. This observation highlights a trade-off between architectural complexity and computational efficiency. While the Chapter 4 framework improves boundary-level agreement and localization robustness, it does so at the cost of increased inference latency.

The reported runtime provides a transparent assessment of computational requirements under the current implementation. Future work may explore optimization strategies, such as model compression, mixed-precision inference, or region pre-screening, to reduce inference time without altering the core methodology.

4.7.1 External Validation on an Independent Dataset

To evaluate generalization beyond the training domain, the proposed framework is assessed on the CAMELYON17 dataset, which is used exclusively for external validation and was not involved in model training or hyperparameter selection. The model trained on CAMELYON16 is applied directly to CAMELYON17 without re-training or fine-tuning, following an evaluation-only protocol.

External evaluation focuses primarily on slide-level classification performance. For each whole-slide image (WSI), a single metastasis likelihood score is computed by aggregating tile-level predictions, and slide-level Area Under the Receiver Operating Characteristic Curve (AUC) is reported as the primary metric. For completeness, lesion-level localization performance is also assessed using the Free-Response Receiver Operating Characteristic (FROC) metric, computed from detected lesion candidates in accordance with standard CAMELYON evaluation practice [21].

Table 4.9 : Comparison of in-domain performance on CAMELYON16 and external validation on CAMELYON17. The CAMELYON17 results reflect generalization performance obtained without re-training or fine-tuning under the same inference pipeline.

Dataset	**AUC (%)**	**FROC (%)**	**Resolution**	**WSIs**
CAMELYON16 (WSI-level split, k-fold)	99.0 ± 0.25	86.2 ± 0.5	10×	40
CAMELYON17 (external validation)	93.5	63.0	10×	N/A

Under this external evaluation setting, a reduction in performance is observed relative

to in-domain results on CAMELYON16. This behavior is consistent with prior CAMELYON studies and reflects the impact of multi-center domain shift arising from differences in staining protocols, scanners, and acquisition conditions [14, 21]. Despite this challenge, the proposed framework maintains strong discriminative performance, indicating robust generalization beyond the training domain.

The quantitative comparison between in-domain evaluation on CAMELYON16 and external validation on CAMELYON17 is summarized in Table 4.9.

4.8 Summary and Discussion

In this chapter, a weakly supervised CAD system has been introduced for simultaneous metastasis classification and lesion-level localization in histopathological WSIs of breast cancer. This approach addresses key challenges associated with limited annotations, computational constraints, and the complex spatial characteristics of metastatic regions. By leveraging GRUs [33] for contextual dependency modeling, ASPP [31] for multi-scale feature aggregation, CBAM [173] for attention-based feature refinement, and skip connections [132] for preserving high-resolution spatial details, the proposed model achieves effective feature representation for metastasis detection and localization. Additionally, the application of a dual-task architecture for classification and tumor ratio regression at the bottleneck and decoder stages enables learning without requiring pixel-level annotations during training. This has been accomplished while utilizing only 1/6 of the annotated CAMELYON16 training dataset at a reduced $10\times$ resolution, under a WSI-level data split with stain normalization, improving computational efficiency.

Experimental results demonstrate strong performance under a consistent and controlled evaluation protocol in both classification and lesion-level localization tasks. The achieved AUC of 0.990 ± 0.25 and FROC score of 0.862 ± 0.5 indicate robust detection capability across WSIs. Furthermore, overlap-based metrics (Dice and IoU) confirm improved boundary-level agreement, supporting accurate delineation of metastatic regions. Compared to previously reported CNN-based approaches, the model achieves competitive

performance while operating within a similar range of computational cost.

External validation on the CAMELYON17 dataset further demonstrates the generalization capability of the proposed framework. Without re-training or fine-tuning, the model maintains competitive classification performance despite the presence of multi-center domain shift. This behavior is consistent with prior studies and highlights the importance of robust feature representation under varying staining and acquisition conditions.

The ablation studies indicate that contextual modeling (via GRUs), multi-scale feature extraction (via ASPP), and feature refinement (via CBAM and skip connections) provide complementary contributions to both lesion detection and boundary-level agreement under a consistent WSI-level evaluation protocol.

Despite its strengths, certain aspects of the model can be further improved. The increased architectural complexity introduces higher computational cost compared to simpler models, reflecting a trade-off between performance and efficiency. In addition, although the model demonstrates good generalization on CAMELYON17, further evaluation on larger and more diverse datasets is required to fully assess robustness across domains.

While the use of 10× magnification significantly improves computational efficiency and enables scalable WSI processing, it inherently limits the availability of fine-grained cytological details. High-magnification views (e.g., 20×–40×) provide additional cellular and nuclear information that can be important for certain diagnostic tasks, as widely discussed in computational pathology literature [101, 86]. As a result, the proposed framework may exhibit reduced sensitivity to very small or low-contrast tumor regions. This reflects a trade-off between efficiency and diagnostic detail, where the current design prioritizes large-scale screening performance.

Future work will focus on addressing these limitations by exploring multi-scale or hierarchical approaches to integrate complementary information across magnification levels. In addition, improving computational efficiency through model compression and efficient inference schemes, as well as extending the framework to multi-class classification and can-

cer grading and investigating self-supervised or semi-supervised learning strategies, may further enhance performance under limited annotation settings.

In conclusion, the proposed CAD system provides a balanced approach to metastasis classification and localization in WSIs under controlled evaluation conditions, combining contextual modeling, multi-scale representation, and attention mechanisms within a weakly supervised setting.

Chapter 5

Summary and Conclusion

Breast cancer remains a critical global health issue, ranking among the most prevalent and fatal diseases affecting women. Early detection and accurate diagnosis are essential for improving patient outcomes, particularly through the identification of metastatic regions, which significantly influence prognosis and treatment decisions. This thesis addresses key challenges in breast cancer histopathology analysis by developing deep learning–based CAD systems for metastasis detection and localization within WSIs. While WSIs provide rich phenotypic information, their large size and the high cost of annotation present significant challenges for practical deployment. The proposed methods aim to balance diagnostic performance with computational efficiency under limited annotation settings.

This research began with a comprehensive review of influential studies published between 2016 and 2024, covering deep learning applications in breast cancer histopathology. The reviewed works were broadly categorized into classification models and detection, localization, and segmentation approaches, with a particular focus on metastasis analysis in whole-slide images (WSIs). In addition, recent advances in foundation models and sequence-based architectures for WSI analysis were discussed, highlighting emerging trends toward large-scale representation learning and improved contextual modeling in computational pathology.

The first framework (Chapter 3) presents an efficient end-to-end architecture for simultaneous WSI classification and lesion-level localization using weak supervision. It operates at a reduced resolution (10×) and avoids the need for pixel-level annotations during training by combining classification and tumor-ratio regression objectives. This design enables computational efficiency and reduced annotation requirements while maintaining competitive performance.

The second framework (Chapter 4) extends this approach by incorporating additional architectural components to improve contextual modeling and boundary-level agreement. Specifically, bidirectional GRUs are integrated to capture long-range dependencies, ASPP is used for multi-scale feature aggregation, and CBAM with skip connections enhances feature refinement and spatial detail preservation. In addition, a more rigorous evaluation protocol is adopted, including WSI-level k-fold cross-validation and stain normalization, to ensure a leakage-resistant and statistically robust assessment.

A comparative analysis of the two frameworks reveals a clear trade-off between efficiency and representational capacity. The Chapter 3 model provides a computationally efficient baseline with strong performance under weak supervision, while the Chapter 4 model achieves improved localization accuracy and boundary-level agreement (as reflected by Dice and IoU metrics, which quantify segmentation overlap) through enhanced contextual and multi-scale modeling. However, these improvements are accompanied by increased computational cost. This comparison highlights the complementary strengths of the two approaches: the first emphasizes efficiency and simplicity, while the second focuses on improved spatial precision and robustness under stricter evaluation conditions.

Experimental results, when evaluated under WSI-level protocols, demonstrate that both frameworks achieve competitive performance in metastasis detection and localization tasks. The results indicate that CNN-based architectures, when carefully designed and evaluated under controlled conditions, can provide strong and competitive baselines for weakly supervised whole-slide image analysis. However, rather than claiming state-of-the-art performance, the findings position the proposed methods within the broader landscape of existing approaches, emphasizing their effectiveness under constrained data and computational settings.

Despite these contributions, several limitations remain. The enhanced architecture in Chapter 4 introduces increased computational overhead, highlighting a trade-off between accuracy and efficiency. Furthermore, while external validation experiments demonstrate reasonable generalization, performance degradation under domain shift indicates that ro-

bustness across diverse clinical settings remains an open challenge. Additionally, the weakly supervised nature of the framework limits the availability of precise pixel-level supervision during training, which may affect boundary accuracy in complex cases. In addition, the use of 10× magnification improves computational efficiency but introduces a trade-off with fine-grained diagnostic detail. This may limit sensitivity to small or subtle metastatic regions that typically require higher magnification (e.g., 20×–40×) for reliable identification.

Future work should focus on several key directions. First, a critical next step is to benchmark the proposed frameworks against modern architectures, including Vision Transformer (ViT)-based models and state-space models such as Mamba-based approaches, under the same controlled evaluation settings. This will provide a clearer understanding of the relative performance of CNN-based models in the current research landscape.

Second, a more rigorous error analysis should be conducted to better understand model limitations. This includes systematically collecting false positive and false negative predictions at both patch and WSI levels, clustering failure cases based on visual characteristics, and analyzing them in collaboration with domain experts. Such analysis can reveal common failure modes, such as sensitivity to staining artifacts or complex tissue structures, and guide targeted model improvements.

Third, future research should explore multimodal data integration, combining histopathology images with complementary information such as pathology reports, genomic data, or other clinical variables. This direction aligns with emerging trends in precision medicine and has the potential to improve diagnostic and prognostic performance beyond image-only models.

Additional directions include improving computational efficiency through model compression and optimized inference strategies, as well as extending the framework to multi-class classification and cancer grading tasks under weak supervision.

Another important direction for future work is uncertainty quantification (UQ) for model predictions. While the proposed frameworks produce probability maps that indicate

metastasis likelihood, these scores are not explicitly calibrated and may not reliably reflect predictive confidence. In high-stakes medical applications, it is critical to distinguish between confident and uncertain predictions to support safe and informed decision-making. Future research will investigate established UQ techniques, such as Monte Carlo dropout and deep ensembles, to estimate prediction uncertainty and improve the reliability and interpretability of model outputs.

In conclusion, this thesis presents two CNN-based frameworks for weakly supervised metastasis detection in whole-slide histopathology images. By balancing efficiency, contextual modeling, and evaluation rigor, the proposed approaches demonstrate strong performance under constrained settings and provide a foundation for further research in computational pathology. These contributions highlight the potential of carefully designed deep learning systems to support and enhance histopathology analysis, while also identifying key challenges and opportunities for future work.

Bibliography

[1] "Breast cancer imaging," http://www.aboutcancer.com/breast_cancer_imaging.htm, 2018, accessed: 10 March 2020.

[2] A. H. Abdulaal, M. Valizadeh, M. C. Amirani, and A. S. Shah, "A self-learning deep neural network for classification of breast histopathological images," *Biomedical Signal Processing and Control*, vol. 87, p. 105418, 2024.

[3] U. R. Acharya, W. L. Ng, K. Rahmat, V. K. Sudarshan, J. E. Koh, J. H. Tan, Y. Hagiwara, A. Gertych, F. Fadzli, C. H. Yeong *et al.*, "Shear wave elastography for characterization of breast lesions: Shearlet transform and local binary pattern histogram techniques," *Computers in Biology and Medicine*, vol. 91, pp. 13–20, 2017.

[4] D. Addo, S. Zhou, K. Sarpong, O. T. Nartey, M. A. Abdullah, C. C. Ukwuoma, and M. A. Al-antari, "A hybrid lightweight breast cancer classification framework using the histopathological images," *Biocybernetics and Biomedical Engineering*, vol. 44, no. 1, pp. 31–54, 2024.

[5] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[6] I. Andersson, D. M. Ikeda, S. Zackrisson, M. Ruschin, T. Svahn, P. Timberg, and A. Tingberg, "Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and birads classification in a population of cancers with subtle mammographic findings," *European radiology*, vol. 18, pp. 2817–2825, 2008.

[7] N. Antropova, H. Abe, and M. L. Giger, "Use of clinical mri maximum intensity projections for improved breast lesion classification with deep convolutional neural networks," *Journal of Medical Imaging*, vol. 5, no. 1, pp. 014 503–014 503, 2018.

[8] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[9] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar, "Bach: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518307941

[10] G. Aresta, T. Araújo, J. Rebelo, S. Rebelo, A. Polónia, and A. Campilho, "Inesc tec - breast cancer histology dataset," https://rdm.inesctec.pt/dataset/nis-2017-003, 2017, accessed: 2025-04-10.

[11] A. Ashurov, S. A. Chelloug, A. Tselykh, M. S. A. Muthanna, A. Muthanna, and M. S. Al-Gaashani, "Improved breast cancer classification through combining transfer learning and attention mechanism," *Life*, vol. 13, no. 9, p. 1945, 2023.

[12] M. Aubreville, C. A. Bertram, M. Veta, R. Klopfleisch, and A. Maier, "Mitosis domain generalization challenge," in *Proceedings of the MICCAI 2021 Mitosis Domain Generalization Challenge*. Springer, 2022.

[13] J. A. Baker and J. Y. Lo, "Breast tomosynthesis: state-of-the-art and review of the literature," *Academic radiology*, vol. 18, no. 10, pp. 1298–1310, 2011.

[14] P. Bandi, O. Geessink, Q. Manson *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2019.

[15] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based

on histology images using convolutional neural networks," *Ieee Access*, vol. 6, pp. 24 680–24 693, 2018.

[16] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," pp. 2440–2445, 2016.

[17] A. S. Becker, M. Mueller, E. Stoffel, M. Marcon, S. Ghafoor, and A. Boss, "Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study," *The British journal of radiology*, vol. 91, no. 1083, p. 20170576, 2018.

[18] K. S. Beevi, M. S. Nair, and G. Bindu, "A multi-classifier system for automatic mitosis detection in breast histopathology images using deep belief networks," *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–11, 2017.

[19] B. E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, and J. A. Van Der Laak, "Automated detection of dcis in whole-slide h&e stained breast histopathology images," *IEEE transactions on medical imaging*, vol. 35, no. 9, pp. 2141–2150, 2016.

[20] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken, N. Karssemeijer, G. Litjens, and J. van der Laak, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *Journal of Medical Imaging*, vol. 4, no. 4, pp. 044 504–044 504, 2017.

[21] B. Bejnordi, M. Veta, P. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. van der Laak, M. Hermsen, Q. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. van Dijk, P. Bult, F. Beca, A. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H.-J. Lin, P.-A. Heng, C. Hass, E. Bruni, Q. Wong, U. Halici, M. Oner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl,

M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venancio, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[22] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[23] C. Boetes and R. M. Mann, "Ductal carcinoma in situ and breast mri," *The Lancet*, vol. 370, no. 9586, pp. 459–460, 2007.

[24] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3279–3298, 2021.

[25] H. C. Burrell, D. M. Sibbering, A. Wilson, S. E. Pinder, A. J. Evans, L. J. Yeoman, C. W. Elston, I. O. Ellis, R. W. Blamey, and J. Robertson, "Screening interval breast cancers: mammographic features and prognosis factors." *Radiology*, vol. 199, no. 3, pp. 811–817, 1996.

[26] M. Byra, T. Sznajder, D. Korzinek, H. Piotrzkowska-Wroblewska, K. Dobruch-Sobczak, A. Nowicki, and K. Marasek, "Impact of ultrasound image reconstruction method on breast lesion classification with deep learning," in *Pattern Recognition and Image Analysis. IbPRIA 2019.* Springer, 2019, pp. 41–52. [Online]. Available: https://doi.org/10.1007/978-3-030-31332-6_4

[27] G. Campanella *et al.*, "A clinical benchmark of public self-supervised pathology foundation models," *Nature Communications*, 2025.

[28] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[29] D. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (froc) data," *Medical Physics*, vol. 16, no. 4, pp. 561–568, 1989.

[30] H. Chen, Q. Dou, X. Wang, J. Qin, and P. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[32] R. J. Chen, T. Ding, M. Y. Lu *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[33] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179/

[34] E. F. Conant, E. F. Beaber, B. L. Sprague, S. D. Herschorn, D. L. Weaver, T. Onega, A. N. Tosteson, A. M. McCarthy, S. P. Poplack, J. S. Haas *et al.*, "Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: a cohort study within the prospr consortium," *Breast cancer research and treatment*, vol. 156, pp. 109–116, 2016.

[35] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and F. González, "Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent," *Scientific Reports*, vol. 7, p. 46450, 2017. [Online]. Available: https://doi.org/10.1038/srep46450

[36] J. de Matos, S. T. M. Ataky, A. de Souza Britto Jr., L. E. S. de Oliveira, and A. L. Koerich, "Machine learning methods for histopathological image analysis: A review," *Electronics*, vol. 10, no. 5, p. 562, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/5/562

[37] T. G. Debelee, F. Schwenker, A. Ibenthal, and D. Yohannes, "Survey of deep learning in breast cancer image analysis," *Evolving Systems*, vol. 11, no. 1, pp. 143–163, 2020.

[38] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[39] T. Ding *et al.*, "A multimodal whole-slide foundation model for pathology," *Nature Medicine*, 2025.

[40] L. Duijm, G. Guit, J. Zaat, A. Koomen, and D. Willebrand, "Sensitivity, specificity and predictive values of breast imaging in the detection of cancer," *British journal of cancer*, vol. 76, no. 3, pp. 377–381, 1997.

[41] Y. Ektefaie, W. Yuan, D. A. Dillon, N. U. Lin, J. A. Golden, I. S. Kohane, and K.-H. Yu, "Integrative multiomics-histopathology analysis for breast cancer classification," *NPJ Breast Cancer*, vol. 7, no. 1, p. 147, 2021.

[42] Z. Fang *et al.*, "Mammil: Multiple instance learning for whole slide images with state space models," *arXiv*, 2024.

[43] N. Farajzadeh, N. Sadeghzadeh, and M. Hashemzadeh, "A fully-convolutional residual encoder-decoder neural network to localize breast cancer on histopathology

images," *Computers in Biology and Medicine*, vol. 147, p. 105698, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482522004814

[44] Y. Faridah, "Digital versus screen film mammography: a clinical comparison," *Biomedical imaging and intervention journal*, vol. 4, no. 4, 2008.

[45] M. F. A. Fauzi, M. F. Jamaluddin, J. T. H. Lee, K. H. Teoh, and L. M. Looi, "Tumor region localization in h&e breast carcinoma images using deep convolutional neural network," in *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, 2018, pp. 61–66.

[46] O. Fourkioti, M. De Vries, and C. Bakal, "Camil: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images," in *International Conference on Learning Representations (ICLR)*, 2024.

[47] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3507–3516, 2021.

[48] A. H. Gandomi and A. H. Alavi, "Krill herd: a new bio-inspired optimization algorithm," *Communications in nonlinear science and numerical simulation*, vol. 17, no. 12, pp. 4831–4845, 2012.

[49] C. Gao, Q. Sun, W. Zhu, L. Zhang, J. Zhang, B. Liu, and J. Zhang, "Transformer based multiple instance learning for wsi breast cancer classification," *Biomedical Signal Processing and Control*, vol. 89, p. 105755, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423011886

[50] B. Gecer, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks," *Pattern recognition*, vol. 84, pp. 345–356, 2018.

[51] G. Gennaro, D. Bernardi, and N. Houssami, "Radiation dose with digital breast

tomosynthesis compared to digital mammography: per-view analysis," *European radiology*, vol. 28, pp. 573–581, 2018.

[52] G. Gennaro, A. Toledano, C. Di Maggio, E. Baldan, E. Bezzon, M. La Grassa, L. Pescarini, I. Polico, A. Proietti, A. Toffoli *et al.*, "Digital breast tomosynthesis versus digital mammography: a clinical performance study," *European radiology*, vol. 20, pp. 1545–1553, 2010.

[53] F. J. Gilbert, L. Tucker, M. G. Gillan, P. Willsher, J. Cooke, K. A. Duncan, M. J. Michell, H. M. Dobson, Y. Y. Lim, T. Suaris *et al.*, "Accuracy of digital breast tomosynthesis for depicting breast cancer subgroups in a uk retrospective reading study (tommy trial)," *Radiology*, vol. 277, no. 3, pp. 697–706, 2015.

[54] U. R. Gogoi, M. K. Bhowmik, D. Bhattacharjee, A. K. Ghosh, and G. Majumdar, "A study and analysis of hybrid intelligent techniques for breast cancer detection using breast thermograms," *Hybrid Soft Computing Approaches: Research and Applications*, pp. 329–359, 2016.

[55] I. Griebsch, J. Brown, C. Boggis, A. Dixon, M. Dixon, D. Easton, R. Eeles, D. Evans, F. J. Gilbert, J. Hawnaur *et al.*, "Cost-effectiveness of screening with contrast enhanced magnetic resonance imaging vs x-ray mammography of women at a high familial risk of breast cancer," *British journal of cancer*, vol. 95, no. 7, pp. 801–810, 2006.

[56] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv*, 2023.

[57] Z. Guo, H. Liu, H. Ni, X. Wang, M. Su, W. Guo, K. Wang, T. Jiang, and Y. Qian, "A fast and refined cancer regions segmentation framework in whole-slide breast pathological images," *Scientific reports*, vol. 9, no. 1, p. 882, 2019.

[58] D. Gur, G. S. Abrams, D. M. Chough, M. A. Ganott, C. M. Hakim, R. L. Perrin, G. Y. Rathfon, J. H. Sumkin, M. L. Zuley, and A. I. Bandos, "Digital breast to-

mosynthesis: observer performance study," *American Journal of Roentgenology*, vol. 193, no. 2, pp. 586–591, 2009.

[59] A. Hadjipanteli, M. Kontos, and A. Constantinidou, "The role of digital breast tomosynthesis in breast cancer screening: a manufacturer-and metrics-specific analysis," *Cancer management and research*, pp. 9277–9296, 2019.

[60] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, and A. Maria Vanegas, "Breast cancer histopathology image classification using an ensemble of deep learning models," *Sensors*, vol. 20, no. 16, p. 4373, 2020.

[61] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Scientific reports*, vol. 7, no. 1, p. 4172, 2017.

[62] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[63] Y. He and X. Li, "Whole-slide-imaging cancer metastases detection and localization with limited tumorous data," in *Proceedings of Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 227, 2024, pp. 877–887.

[64] M. A. Helvie, "Digital mammography imaging: breast tomosynthesis and advanced applications," *Radiologic Clinics*, vol. 48, no. 5, pp. 917–929, 2010.

[65] I. Hirra, M. Ahmad, A. Hussain, M. U. Ashraf, I. A. Saeed, S. F. Qadri, A. M. Alghamdi, and A. S. Alfakeeh, "Breast cancer classification from histopathological images using patch-based deep learning modeling," *IEEE Access*, vol. 9, pp. 24 273–24 287, 2021.

[66] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[67] R. J. Hooley, M. A. Durand, and L. E. Philpotts, "Advances in digital breast tomosynthesis," *American Journal of Roentgenology*, vol. 208, no. 2, pp. 256–266, 2017.

[68] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2424–2433.

[69] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[70] J. Huang *et al.*, "Transformer optimization with meta learning on pathology images for breast cancer lymph node micrometastasis," *npj Digital Medicine*, vol. 8, 2025.

[71] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[72] M. Jahanifar, A. Shephard, N. Zamanitajeddin, S. Graham, S. E. A. Raza, F. Minhas, and N. Rajpoot, "Mitosis detection, fast and slow: robust and efficient detection of mitotic figures," *Medical Image Analysis*, vol. 94, p. 103132, 2024.

[73] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical imaging*, vol. 37, no. 3, pp. 420–426, 2013.

[74] S. Javed, A. Mahmood, T. Qaiser, N. Werghi, and N. Rajpoot, "Unsupervised mutual transformer learning for multi-gigapixel whole slide image classification," *Medical Image Analysis*, vol. 96, p. 103203, 2024.

[75] O. Jimenez-del Toro, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, M. Rousson, H. Müller, and M. Atzori, "Analysis of histopathology images: From traditional machine learning to deep learning," in *Biomedical texture analysis*. Elsevier, 2017, pp. 281–314.

[76] A. Juhong, B. Li, C.-Y. Yao, C.-W. Yang, D. W. Agnew, Y. L. Lei, X. Huang, W. Piyawattanametha, and Z. Qiu, "Super-resolution and segmentation deep learning for breast cancer histopathology image analysis," *Biomedical optics express*, vol. 14, no. 1, pp. 18–36, 2023.

[77] H. Kang, Q. Xu, D. Chen, S. Ren, H. Xie, L. Wang, Y. Gao, M. Gong, and X. Chen, "Assessing the performance of fully supervised and weakly supervised learning in breast cancer histopathology," *Expert Systems with Applications*, vol. 237, p. 121575, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423020778

[78] 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," *arXiv preprint arXiv:1909.11870*, 2019, preprint, not peer-reviewed. No published version available as of April 2025.

[79] J. N. Kather, "Histological images for msi vs. mss classification in gastrointestinal cancer, ffpe samples," https://doi.org/10.5281/zenodo.2530835, 2019, accessed: 2025-04-10.

[80] K. M. Kelly, J. Dean, W. S. Comulada, and S.-J. Lee, "Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts," *European radiology*, vol. 20, pp. 734–742, 2010.

[81] M.-A. Khalil, Y.-C. Lee, H.-C. Lien, Y.-M. Jeng, and C.-W. Wang, "Fast segmentation of metastatic foci in h&e whole-slide images for breast cancer diagnosis," *Diagnostics*, vol. 12, no. 4, p. 990, 2022.

[82] S. Khaliliboroujeni, X. He, W. Jia, and S. Amirgholipour, "End-to-end metastasis detection of breast cancer from histopathology whole slide images," *Computerized Medical Imaging and Graphics*, vol. 102, p. 102136, 2022.

[83] H.-E. Kim and S. Hwang, "Deconvolutional feature stacking for weakly-supervised

semantic segmentation," *arXiv preprint arXiv:1602.04984*, 2016, preprint, not peer-reviewed. No published version available as of April 2025.

[84] J. Kim, A. Harper, V. McCormack *et al.*, "Global patterns and trends in breast cancer incidence and mortality across 185 countries," *Nature Medicine*, 2025. [Online]. Available: https://doi.org/10.1038/s41591-025-03502-3

[85] Y. J. Kim, H. Jang, K. Lee, S. Park, S.-G. Min, C. Hong, J. H. Park, K. Lee, J. Kim, W. Hong *et al.*, "Paip 2019: Liver cancer segmentation challenge," *Medical image analysis*, vol. 67, p. 101854, 2021.

[86] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and Structural Biotechnology Journal*, vol. 16, pp. 34–42, 2018.

[87] B. Kong, X. Wang, Z. Li, Q. Song, and S. Zhang, "Cancer metastasis detection via spatially structured deep network," in *International Conference on Information Processing in Medical Imaging*, 2017, pp. 236–248.

[88] D. B. Kopans, "Digital breast tomosynthesis from concept to clinical care," *American Journal of Roentgenology*, vol. 202, no. 2, pp. 299–308, 2014.

[89] C. K. Kuhl, "The changing world of breast cancer: a radiologist's perspective," *Investigative radiology*, vol. 50, no. 9, pp. 615–628, 2015.

[90] K. Lång, I. Andersson, A. Rosso, A. Tingberg, P. Timberg, and S. Zackrisson, "Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the malmö breast tomosynthesis screening trial, a population-based study," *European radiology*, vol. 26, pp. 184–190, 2016.

[91] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification," *Medical Image Analysis*, vol. 69, p. 101979, 2021.

[92] ——, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 318–14 328.

[93] C. Li, X. Wang, W. Liu, L. J. Latecki, B. Wang, and J. Huang, "Weakly supervised mitosis detection in breast histopathology images using concentric loss," *Medical image analysis*, vol. 53, pp. 165–178, 2019.

[94] J. Li, W. Li, A. Sisk, H. Ye, W. D. Wallace, W. Speier, and C. W. Arnold, "A multi-resolution model for histopathology image classification and localization with multiple instance learning," *Computers in biology and medicine*, vol. 131, p. 104253, 2021.

[95] J. Li, A. Osseyran, R. Hekster, S. Rudinac, V. Codreanu, and D. Podareanu, "Improving the speed and quality of cancer segmentation using lower resolution pathology images," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 11 999–12 015, 2024.

[96] K. Li, Z. Qian, Y. Han, I. Eric, C. Chang, B. Wei, M. Lai, J. Liao, Y. Fan, and Y. Xu, "Weakly supervised histopathology image segmentation with self-attention," *Medical Image Analysis*, vol. 86, p. 102791, 2023.

[97] Y. Li, Y. Xue, L. Li, X. Zhang, and X. Qian, "Domain adaptive box-supervised instance segmentation network for mitosis detection," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2469–2485, 2022.

[98] H. Lin, H. Chen, Q. Dou, L. Wang, J. Qin, and P.-A. Heng, "ScanNet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 539–546.

[99] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P.-A. Heng, "Fast ScanNet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis

detection," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1948–1958, 2019.

[100] S.-P. Lin and J. J. Brown, "Mr contrast agents: physical and pharmacologic basics," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 25, no. 5, pp. 884–899, 2007.

[101] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[102] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017, preprint, not peer-reviewed. No published version available as of April 2025.

[103] A. P. Lourenco, M. Barry-Brooks, G. L. Baird, A. Tuttle, and M. B. Mainiero, "Changes in recall type and patient treatment following implementation of screening digital breast tomosynthesis," *Radiology*, vol. 274, no. 2, pp. 337–342, 2015.

[104] M. Y. Lu *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 863–874, 2024.

[105] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[106] Z. Lu, K. Tang, Y. Wu, X. Zhang, Z. An, X. Zhu, Q. Feng, and Y. Zhao, "Breastdluseg: A coarse-to-fine framework for segmentation of breast terminal duct lobular units on histopathological whole-slide images," *Computerized Medical Imaging and Graphics*, vol. 118, p. 102432, 2024.

[107] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for

quantitative analysis," *IEEE International Symposium on Biomedical Imaging*, pp. 1107–1110, 2009.

[108] T. Mahmood, M. Arsalan, M. Owais, M. B. Lee, and K. R. Park, "Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs," *Journal of clinical medicine*, vol. 9, no. 3, p. 749, 2020.

[109] S. Mall, S. Lewis, P. Brennan, J. Noakes, and C. Mello-Thoms, "The role of digital breast tomosynthesis in the breast assessment clinic: a review," *Journal of medical radiation sciences*, vol. 64, no. 3, pp. 203–211, 2017.

[110] M. T. McCann, J. A. Ozolek, C. A. Castro, B. Parvin, and J. Kovacevic, "Automated histology analysis: Opportunities for signal processing," *IEEE Signal Processing Magazine*, vol. 32, no. 1, pp. 78–87, 2014.

[111] R. Mehra *et al.*, "Breast cancer histology images classification: Training from scratch or transfer learning?" *Ict Express*, vol. 4, no. 4, pp. 247–254, 2018.

[112] Y. Mina, K. Adam, and Y. S. Ching, "Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning," *Computers in Biology and Medicine*, vol. 96, pp. 283–293, 2018. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2018.04.004

[113] MITOS-ATYPIA-14 Organizers, "Mitos-atypia-14 grand challenge dataset," https://mitos-atypia-14.grand-challenge.org/Dataset/, 2014, accessed: 2025-04-10.

[114] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.

[115] G. Murtaza, L. Shuib, A. W. Abdul Wahab, G. Mujtaba, G. Mujtaba, H. F. Nweke, M. A. Al-garadi, F. Zulfiqar, G. Raza, and N. A. Azmi, "Deep learning-based breast

cancer classification through medical imaging modalities: state of the art and research challenges," *Artificial Intelligence Review*, vol. 53, pp. 1655–1720, 2020.

[116] G. Murtaza, L. Shuib, G. Mujtaba, and G. Raza, "Breast cancer multi-classification through deep neural network and hierarchical classification approach," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 15 481–15 511, 2020.

[117] M. Nawaz, A. Sewissy, and T. Soliman, "Multi-class breast cancer classification using deep learning convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 9, pp. 316–322, 2018.

[118] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[119] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[120] S. Ortiz, I. Rojas-Valenzuela, F. Rojas, O. Valenzuela, L. J. Herrera, and I. Rojas, "Novel methodology for detecting and localizing cancer area in histopathological images based on overlapping patches," *Computers in Biology and Medicine*, vol. 168, p. 107713, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523011782

[121] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[122] S. K. Patterson and M. A. Roubidoux, "Update on new technologies in digital mammography," *International journal of women's health*, pp. 781–788, 2014.

[123] M. Pereira, "Breakhis-breast cancer histopathological database," *Mendeley Data*, vol. 1, 2023.

[124] L. E. Philpotts and M. A. Durand, "Good news for older women undergoing screening mammography with digital breast tomosynthesis," pp. 591–593, 2019.

[125] S. P. Poplack, T. D. Tosteson, C. A. Kogel, and H. M. Nagy, "Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography," *American Journal of Roentgenology*, vol. 189, no. 3, pp. 616–623, 2007.

[126] B. M. Priego-Torres, D. Sanchez-Morillo, M. A. Fernandez-Granero, and M. Garcia-Rojo, "Automatic segmentation of whole-slide h&e stained breast histopathology images using a deep convolutional neural network architecture," *Expert Systems With Applications*, vol. 151, p. 113387, 2020.

[127] L. Qiu, L. Zhao, R. Hou, W. Zhao, S. Zhang, Z. Lin, H. Teng, and J. Zhao, "Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features," *Computerized Medical Imaging and Graphics*, vol. 104, p. 102176, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089561112200146X

[128] E. A. Rafferty, M. A. Durand, E. F. Conant, D. S. Copit, S. M. Friedewald, D. M. Plecha, and D. P. Miller, "Breast cancer screening using tomosynthesis and digital mammography in dense and nondense breasts," *Jama*, vol. 315, no. 16, pp. 1784–1786, 2016.

[129] E. A. Rafferty, J. M. Park, L. E. Philpotts, S. P. Poplack, J. H. Sumkin, E. F. Halpern, and L. T. Niklason, "Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial," *Radiology*, vol. 266, no. 1, pp. 104–113, 2013.

[130] J. Rahnfeld *et al.*, "A comparative study of explainability methods for whole slide classification of lymph node metastases using vision transformers," *PLOS Digital Health*, 2025.

[131] K. Rangarajan, S. Hari, S. Thulkar, S. Sharma, A. Srivastava, and R. Parshad, "Characterization of lesions in dense breasts: Does tomosynthesis help?" *Indian Journal of Radiology and Imaging*, vol. 26, no. 02, pp. 210–215, 2016.

[132] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[133] Y. Rui, L. Pei, and J. Luping, "Collaborative learning with curriculum loss for accurate and interpretableweakly supervised whole-slide image classification collaborative learning with curriculum loss for whole-slide image classification," in *Proceedings of 2024 International Conference on Machine Learning and Intelligent Computing*, ser. Proceedings of Machine Learning Research, Z. Nianyin and R. B. Pachori, Eds., vol. 245. PMLR, 26–28 Apr 2024, pp. 33–39. [Online]. Available: https://proceedings.mlr.press/v245/rui24a.html

[134] M. Saha and C. Chakraborty, "Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2189–2200, 2018.

[135] M. Saha, C. Chakraborty, and D. Racoceanu, "Efficient deep learning model for mitosis detection using breast histopathology images," *Computerized Medical Imaging and Graphics*, vol. 64, pp. 29–40, 2018.

[136] H. Sasabe, T. Wada, M. Hosoda, H. Okawa, M. Hara, A. Yamada, and A. F. Garito, "Photonics and organic nanostructures," in *Nonlinear Optical Properties of Organic Materials III*, vol. 1337. SPIE, 1990, pp. 62–72.

[137] Y. Schirris, E. Gavves, I. Nederlof, H. M. Horlings, and J. Teuwen, "Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer," *Medical Image Analysis*, vol. 79, p. 102464, 2022.

[138] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[139] R. Schmitz, F. Madesta, M. Nielsen, J. Krause, S. Steurer, R. Werner, and T. Rösch, "Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture," *Medical image analysis*, vol. 70, p. 101996, 2021.

[140] M. Sebai, T. Wang, and S. A. Al-Fadhli, "PartMitosis: a partially supervised deep learning framework for mitosis detection in breast cancer histopathology images," *IEEE Access*, vol. 8, pp. 45 133–45 147, 2020.

[141] Z. Shao *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[142] R. Shi, X. Li, and H. Gao, "A novel deep learning model csam-resnet based on attention mechanism and residual learning for breast histopathology image classification," in *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2024, pp. 3042–3047.

[143] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim, and K. M. Lee, "Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images," *IEEE transactions on medical imaging*, vol. 38, no. 3, pp. 762–774, 2018.

[144] M. S. H. Shovon, M. F. Mridha, K. M. Hasib, S. Alfarhood, M. Safran, and D. Che, "Addressing uncertainty in imbalanced histopathology image classification of her2 breast cancer: An interpretable ensemble approach with threshold filtered single instance evaluation (sie)," *IEEE Access*, vol. 11, pp. 122 238–122 251, 2023.

[145] A. Singha and M. K. Bhowmik, "Alexsegnet: an accurate nuclei segmentation deep learning model in microscopic images for diagnosis of cancer," *Multimedia tools and Applications*, vol. 82, no. 13, pp. 20 431–20 452, 2023.

[146] P. Skaane, "Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review," *Acta radiologica*, vol. 50, no. 1, pp. 3–14, 2009.

[147] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[148] M. M. Srikantamurthy, V. S. Rallabandi, D. B. Dudekula, S. Natarajan, and J. Park, "Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid cnn-lstm based transfer learning," *BMC Medical Imaging*, vol. 23, no. 1, p. 19, 2023.

[149] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[150] H. Su, D. Zhao, A. A. Heidari, L. Liu, X. Zhang, M. Mafarja, and H. Chen, "RIME: A physics-based optimization," *Neurocomputing*, vol. 532, pp. 183–214, 2023.

[151] V. K. Sudarshan, M. R. K. Mookiah, U. R. Acharya, V. Chandran, F. Molinari, H. Fujita, and K. H. Ng, "Application of wavelet techniques for cancer diagnosis using ultrasound images: a review," *Computers in biology and medicine*, vol. 69, pp. 97–111, 2016.

[152] P. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, and P. Honeine, "Multiple instance learning for histopathological breast cancer image classification," *Expert Systems with Applications*, vol. 117, pp. 103–111, 2019.

[153] S. Taheri and Z. Golrizkhatami, "Magnification-specific and magnification-independent classification of breast cancer histopathological image using deep learning approaches," *Signal, Image and Video Processing*, vol. 17, no. 2, pp. 583–591, 2023.

[154] S. Tessa, J. Keith, and F. Keith, "The difference between an mri and ct scan," 2018.

[155] The Cancer Genome Atlas Research Network, "Comprehensive genomic characterization of diffuse large b-cell lymphoma," *Nature*, vol. 490, pp. 116–120, 2013.

[156] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemporary Oncology (Poznan, Poland)*, vol. 19, no. 1A, pp. A68–A77, 2015.

[157] P. Tourniaire, M. Ilie, P. Hofman, N. Ayache, and H. Delingette, "Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images," *Medical Image Analysis*, vol. 85, p. 102763, 2023.

[158] P.-H. Tsui, Y.-Y. Liao, C.-C. Chang, W.-H. Kuo, K.-J. Chang, and C.-K. Yeh, "Classification of benign and malignant breast tumors by 2-d analysis based on contour description and scatterer characterization," *IEEE transactions on medical imaging*, vol. 29, no. 2, pp. 513–522, 2010.

[159] C. C. Ukwuoma, M. A. Hossain, J. K. Jackson, G. U. Nneji, H. N. Monday, and Z. Qin, "Multi-classification of breast cancer lesions in histopathological images using deep_pachi: Multiple self-attention head," *Diagnostics*, vol. 12, no. 5, p. 1152, 2022.

[160] R. Valieris, L. Martins, A. Defelicibus *et al.*, "Weakly-supervised deep learning models enable her2-low prediction from h&e stained slides," *Breast Cancer Research*, vol. 26, p. 124, 2024. [Online]. Available: https://doi.org/10.1186/s13058-024-01863-0

[161] M. Van Rijthoven, M. Balkenhol, K. Siliņa, J. Van Der Laak, and F. Ciompi, "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Medical image analysis*, vol. 68, p. 101890, 2021.

[162] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[163] M. Veta, P.-A. Heng, N. Stathonikos, B. E. Bejnordi, W. Bulten, G. Litjens, P. J. van Diest, and J. A. van der Laak, "Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge," *Medical Image Analysis*, vol. 54, pp. 111–121, 2019.

[164] D. M. Vo, N.-Q. Nguyen, and S.-W. Lee, "Classification of breast cancer histology images using incremental boosting convolution networks," *Information Sciences*, vol. 482, pp. 123–138, 2019.

[165] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[166] E. Vorontsov *et al.*, "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature Medicine*, vol. 30, no. 10, pp. 2924–2935, 2024.

[167] N. Wahab, A. Khan, and Y. S. Lee, "Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images," *Microscopy*, vol. 68, no. 3, pp. 216–233, 2019.

[168] M. G. Wallis, E. Moa, F. Zanca, K. Leifland, and M. Danielsson, "Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution x-ray imaging observer study," *Radiology*, vol. 262, no. 3, pp. 788–796, 2012.

[169] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016, preprint, not peer-reviewed. No published version available as of April 2025.

[170] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical image analysis*, vol. 81, p. 102559, 2022.

[171] S. Wazir and M. M. Fraz, "Histoseg: Quick attention with multi-loss function for

multi-structure segmentation in digital histology images," in *2022 12th International Conference on Pattern Recognition Systems (ICPRS)*. IEEE, 2022, pp. 1–7.

[172] F. Wilm, C. Marzahl, K. Breininger, and M. Aubreville, "Domain adversarial retinanet as a reference algorithm for the mitosis domain generalization challenge," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 5–13.

[173] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[174] L. Xia, Z. Qu, J. An, and Z. Gao, "A weakly supervised method with colorization for nuclei segmentation using point annotations," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[175] M. Xiao, Y. Li, X. Yan, M. Gao, and W. Wang, "Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example," in *Proceedings of the 2024 7th International Conference on Machine Vision and Applications*, 2024, pp. 145–149.

[176] Y. Xie, J. Long, J. Hou, D. Chen, and G. Guan, "Weakly supervised pathological whole slide image classification based on contrastive learning," *Multimedia Tools and Applications*, vol. 83, no. 21, pp. 60 809–60 831, 2024.

[177] Y. Xie, B. Yang, Q. Guan, J. Zhang, Q. Wu, and Y. Xia, "Attention mechanisms in medical image segmentation: A survey," *arXiv preprint arXiv:2305.17937*, 2023.

[178] J. Xing, A. A. Heidari, H. Chen, and H. Zhao, "WHRIME: A weight-based recursive hierarchical RIME optimizer for breast cancer histopathology image segmentation," *Displays*, vol. 82, p. 102648, 2024.

[179] R. Yan, F. Ren, Z. Wang, L. Wang, T. Zhang, Y. Liu, X. Rao, C. Zheng, and

F. Zhang, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, vol. 173, pp. 52–60, 2020.

[180] H. Yao, Z. Wang, M. Zhang, W. Zhu, L. Zhang, L. He, and J. Zhang, "Second-order self-supervised learning for breast cancer classification," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[181] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, p. 101789, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841520301535

[182] Q. Zeng *et al.*, "Mambamil+: Modeling long-term contextual patterns for gigapixel whole slide image," *arXiv*, 2025.

[183] E. T. Zewdie, A. W. Tessema, and G. L. Simegn, "Classification of breast cancer types, sub-types and grade from histopathological images using deep learning technique," *Health and Technology*, vol. 11, no. 6, pp. 1277–1290, 2021.

[184] J. Zhang, A. Saha, Z. Zhu, and M. A. Mazurowski, "Breast tumor segmentation in dce-mri using fully convolutional networks with an application in radiogenomics," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. SPIE, 2018, pp. 192–196.

[185] M. Zhang, B. Dong, and Q. Li, "Ms-gwnn: Multi-scale graph wavelet neural network for breast cancer diagnosis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5.

[186] X. Zhang, X. Zhu, K. Tang, Y. Zhao, Z. Lu, and Q. Feng, "DDTNet: A dense dual-task network for tumor-infiltrating lymphocyte detection and segmentation in histopathological images of breast cancer," *Medical image analysis*, vol. 78, p. 102415, 2022.

[187] X. Zhang, C. Liu, T. Li, and Y. Zhou, "The whole slide breast histopathology image detection based on a fused model and heatmaps," *Biomedical Signal Processing and Control*, vol. 82, p. 104532, 2023.

[188] T. Zheng, K. Jiang, Y. Xiao, S. Zhao, and H. Yao, "M3amba: Memory mamba is all you need for whole slide image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 15 601–15 610.

[189] Y. Zheng, Z. Jiang, F. Xie, H. Zhang, Y. Ma, H. Shi, and Y. Zhao, "Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification," *Pattern Recognition*, vol. 71, pp. 14–25, 2017.

[190] Y. Zhou, C. Zhang, and S. Gao, "Breast cancer classification from histopathological images using resolution adaptive network," *IEEE Access*, vol. 10, pp. 35 977–35 991, 2022.

[191] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact cnns," *BMC medical informatics and decision making*, vol. 19, pp. 1–17, 2019.

[192] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. Heckbert, Ed. Academic Press Professional, Inc., 1994, pp. 474–485.

[193] M. Zweig and G. Campbell, "Receiver-operating characteristic (roc) plots - a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, pp. 561–577, 1993.