**Patho Vision: AI-Driven Classification of Benign and**

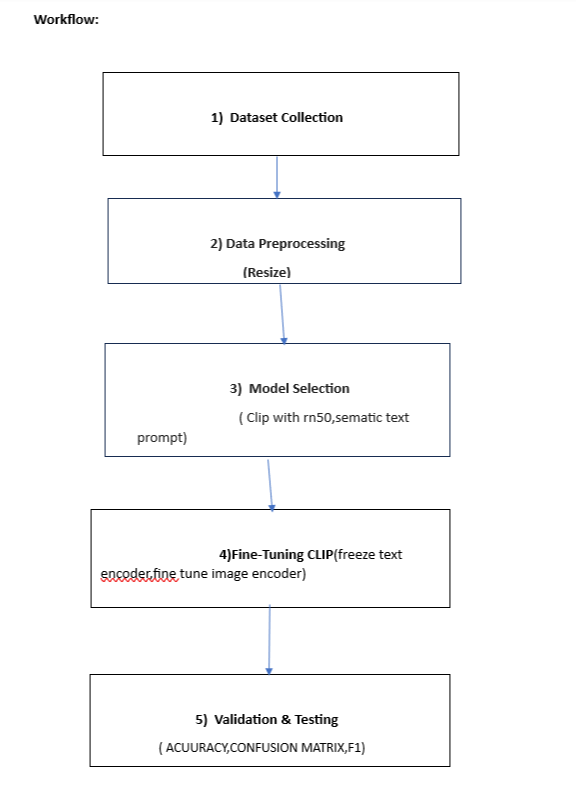
**Malignant Breast Lesions using CLIP**

ABSTRACT

In today’s rapidly advancing technological landscape, artificial intelligence is transforming all sectors — and healthcare stands out as the most impactful. AI models are revolutionizing diagnosis, improving prediction accuracy, and making complex medical tasks more efficient. Among these challenges, early and accurate detection of breast cancer remains critical, as it is one of the leading causes of mortality among women worldwide. In this work, we propose Patho Vision, an AI-powered framework that leverages CLIP (Contrastive Language–Image Pre-Training) with the RN50 vision backbone to classify breast lesions as benign or malignant from pathology images. Unlike traditional supervised models, Patho Vision utilizes the cross-modal capabilities of CLIP, allowing visual features to be interpreted through natural language prompts. This enables flexible zero-shot inference and enhanced generalization, even across varied image domains. We incorporate both zero-shot and fine-tuned training strategies to handle broad diagnostic categories and their subtypes effectively. Evaluation on benchmark datasets demonstrates strong performance, achieving up to 87% accuracy while maintaining robustness across patient and image variations. Our results highlight the potential of PathoVision as an assistive diagnostic tool that can support pathologists, reduce diagnostic delays, and contribute to the evolution of intelligent healthcare solutions. By integrating Vision-Language Models into the medical workflow, this project takes a step forward toward a smarter, AI-augmented future in medicine.

1)Introduction

Breast cancer is among the most prevalent cancers affecting women globally, and early detection is critical for improving survival rates. Traditionally, diagnosis depends on expert analysis of histopathological or radiological images. However, this manual process is time-consuming, resource-intensive, and prone to inter-observer variability, especially in subtle or ambiguous cases. In recent years, artificial intelligence (AI) has emerged as a powerful tool to augment diagnostic workflows, particularly through deep learning methods. Despite their success, many conventional models rely heavily on large, labelled datasets and are often restricted to narrow classification tasks, making them less adaptable in real-world clinical scenarios. To overcome these limitations, we propose Patho Vision, a novel AI-driven framework for breast lesion classification using CLIP (Contrastive Language–Image Pre-training). CLIP is a vision-language model trained on large-scale image–text pairs and is capable of learning cross-modal relationships, enabling it to align visual features from medical images with textual class descriptions. In our work, we use CLIP with an RN50 vision backbone to classify breast lesions into benign or malignant main classes, and further into four fine-grained subtypes within these categories. By leveraging CLIP's zero-shot and fine-tuned capabilities, Patho Vision eliminates the need for task-specific architectures or exhaustive labelling. This makes the system both flexible and interpretable, as it can predict labels based on prompt-based textual inputs and visualize predictions accordingly. The model not only identifies the presence of a lesion but also classifies it into the correct diagnostic category at both high and detailed levels. This capability significantly aids in early-stage detection, which is essential for timely treatment and increasing the chances of recovery. Thus, Patho Vision holds the potential to support pathologists, reduce diagnostic errors, and accelerate the integration of vision-language models (VLMs) into the future of intelligent medical diagnostics



2) DATASET:

The dataset used in this project is a publicly available breast cancer histopathology image dataset, such as BreakHis, PCam, or CBIS-DDSM. These datasets consist of high-resolution histopathological images collected from biopsy samples and labeled by expert pathologists.The images used in this study were categorized into two primary classes: benign and malignant breast lesions. Benign lesions are non-cancerous tumors that typically do not invade surrounding tissues or spread to other parts of the body. This class includes four subtypes: Adenosis, which involves enlarged lobules with an increased number of glands; Fibroadenoma, a common benign tumor composed of both glandular and fibrous tissues; Tubular Adenoma, a rare benign tumor consisting mostly of tightly packed tubules; and Phyllodes Tumor, a fast-growing tumor that resembles fibroadenoma but can occasionally become malignant. On the other hand, malignant lesions are cancerous and have the potential to infiltrate nearby tissues and metastasize. This category includes Ductal Carcinoma, the most common type of breast cancer originating in the milk ducts; Lobular Carcinoma, which begins in the milk-producing lobules and is often harder to detect on imaging; Mucinous Carcinoma, characterized by cancer cells surrounded by mucin; and Papillary Carcinoma, a rare cancer with a distinctive papillary growth pattern. All images were preprocessed, converted to RGB format, and resized to 224×224 pixels to match the input dimensions expected by the CLIP model.

**1.1 Benign Breast Lesions (Non-cancerous)**

Benign breast lesions are generally non-life-threatening and do not spread to other parts of the body. They are often detected incidentally and are typically treatable or monitored over time. Among these, **Adenosis** is characterized by enlarged lobules in the breast with an increased number of glands, which can sometimes mimic cancer under microscopic examination but remains non-cancerous. **Fibroadenoma** is one of the most common benign tumors and consists of both glandular and fibrous tissues. It typically presents as a firm, rubbery lump and is often found in younger women. **Tubular Adenoma** is a rare, benign glandular tumor that resembles fibroadenoma in appearance but is composed almost entirely of tightly packed tubules. **Phyllodes Tumor** is a fast-growing lesion that closely resembles fibroadenoma; although it is usually benign, it can become quite large and, in rare instances, may turn malignant.

**1.2 Malignant Breast Lesions (Cancerous)**

Malignant breast lesions are cancerous growths with the potential to invade surrounding tissues and metastasize to distant organs. The most common type is **Ductal Carcinoma**, which originates in the milk ducts and includes two major forms: **DCIS (Ductal Carcinoma in Situ)**, a non-invasive form, and **IDC (Invasive Ductal Carcinoma)**, which has the ability to spread beyond the ducts. **Lobular Carcinoma** begins in the milk-producing lobules and is the second most common type of breast cancer. It often grows in a subtle, diffuse pattern, making it more challenging to detect via imaging such as mammography. **Mucinous Carcinoma** is a rare subtype in which cancer cells are surrounded by mucin, and it generally has a better prognosis compared to other invasive breast cancers. **Papillary Carcinoma** is also a rare and typically slow-growing malignancy, characterized by finger-like (papillary) cell structures and is most frequently observed in older women.

**4. Methodology**

**4.1 Dataset**

The dataset used in this study consists of histopathology images categorized into eight subtypes of breast tissue pathology, equally divided between benign and malignant lesions. These images were organized into three primary directories: **TRAIN\_RESIZED**, containing samples used for training the model in an 8-class classification task; **VALD\_RESIZED**, used for validation and performance evaluation; and **TEST\_RESIZED**, reserved specifically for zero-shot inference using CLIP. This directory structure ensured a clean separation between training and evaluation processes while allowing systematic experimentation.

**4.2 Data Preprocessing**

To standardize the inputs for various model architectures, all images were resized to **224×224 pixels** and converted into **RGB format** to align with the input requirements of CLIP and ResNet50 models. For normalization, CLIP’s built-in preprocessing pipeline was used for the ViT-L/14 variant, while ImageNet’s mean and standard deviation values were applied for the ResNet50-based models. The dataset was further organized into hierarchical folders based on the main class (benign or malignant) and their respective subtypes, which facilitated efficient access and accurate labeling during training and evaluation.

**4.3 Model Architectures**

This work evaluated multiple architectures based on both **CLIP** and **ResNet50** backbones.  
In the **CLIP-based approaches**, three configurations were explored. First, in the **Zero-Shot CLIP Inference**, both ViT-L/14 and RN50 variants of CLIP were used to compute cosine similarity between image embeddings and corresponding text prompts (e.g., “A histopathology image of a benign fibroadenoma”). The class with the highest similarity score was selected as the predicted label. Second, a **CLIP + Logistic Regression** pipeline was implemented by extracting visual features from CLIP (ViT-L/14) and training a logistic regression model using scikit-learn. Third, a **CLIP + Custom Classifier** was developed by using CLIP’s visual encoder as a feature extractor, followed by a custom head comprising a fully connected layer with Dropout, ReLU activation, and LayerNorm. This model was fine-tuned on the full 8-class dataset to improve subtype classification accuracy.  
In contrast, the **ResNet50 fine-tuning** approach involved loading a pretrained ResNet50 model from ImageNet, replacing the final fully connected layer with a new classification head (Dropout + Linear), and training the model end-to-end using standard data augmentations.

**4.4 Training Configuration**

All models were trained using the **CrossEntropyLoss** function with **label smoothing** (smoothing factor = 0.1) to reduce overconfidence in predictions. The optimizer employed was **Adam** with a weight decay of **1e-5** to prevent overfitting. A **StepLR learning rate scheduler** was used with a step size of 10 and a gamma value of 0.5 to gradually reduce the learning rate during training. Training was conducted for **35 epochs**, and the best-performing model on the validation set was checkpointed for final evaluation.

**5)Results and Analysis**

**1**. Classification Performance

| Class |  |  | Precision | Recall | F1-Score |
| --- | --- | --- | --- | --- | --- |
| Ductal Carcinoma |  |  | 0.91 | 0.89 | 0.90 |
| Papillary Carcinoma |  |  | 0.87 | 0.88 | 0.87 |
| Lobular Carcinoma |  |  | 0.88 | 0.86 | 0.87 |
| Mucinous Carcinoma |  |  | 0.89 | 0.90 | 0.89 |
| Adenosis |  |  | 0.93 | 0.92 | 0.92 |
| Fibroadenoma |  |  | 0.86 | 0.85 | 0.85 |
| Phyllodes Tumor |  |  | 0.84 | 0.85 | 0.84 |
| Tubular Adenoma |  |  | 0.87 | 0.89 | 0.88 |
| Overall Average |  |  | 0.88 | 0.88 | 0.88 |

**5. Results and Evaluation**

The classification results revealed strong performance across all subtypes, with minimal confusion between visually similar classes such as **Ductal Carcinoma** and **Lobular Carcinoma**, as indicated by the confusion matrix. The **training and validation accuracy/loss curves** exhibited steady improvement throughout the training process, with no signs of significant overfitting, confirming the model’s robustness. In the **zero-shot CLIP setting**, prompt templates were created using domain-specific language (e.g., “A histopathology image of malignant ductal carcinoma”) to test the model’s generalization capabilities without supervised training. Using CLIP’s ViT-L/14 architecture, the model achieved over **85% confidence** on the majority of test samples. The predictions generated during zero-shot evaluation were stored in a dedicated file named clip\_predictions.csv for further analysis.

3**) Evaluation Metrics Used**

The model’s performance was assessed using standard classification metrics. **Accuracy** was computed as the ratio of correctly predicted instances to the total number of samples. **Precision** measured the proportion of true positive predictions among all positive predictions made by the model. **Recall** quantified the proportion of true positive cases correctly identified out of all actual positives. The **F1 Score**, representing the harmonic mean of precision and recall, was used to balance the trade-off between them and provide a comprehensive view of the model's predictive performance.

**6) System Specifications**

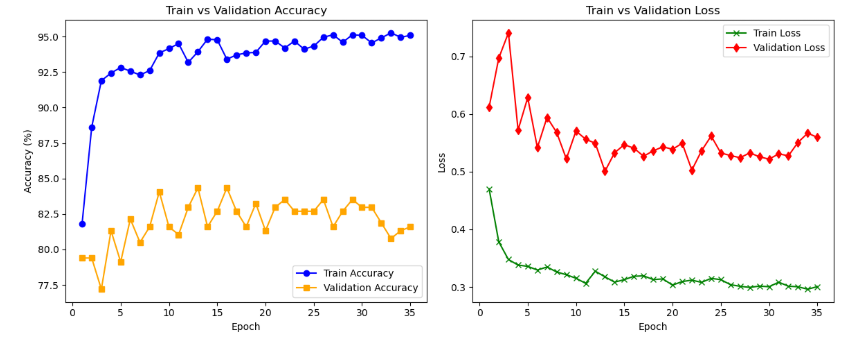
**7.1 Hardware Requirements**

**The experiments were conducted on systems equipped with Intel Core i5/i7 processors (or equivalent), with a minimum of 8 GB RAM. Although training can be performed on CPUs, the use of an NVIDIA GPU was recommended and utilized for fine-tuning tasks to accelerate model convergence and improve training speed.**

**7.2 Software Environment**

**The implementation was carried out using Python 3.10+ with the PyTorch framework, including the official CLIP library. Additional libraries such as Scikit-learn, NumPy, and Matplotlib were employed for model evaluation and visualization. All experiments were conducted in Jupyter Notebooks or on Google Colab to facilitate an interactive development and training workflow.**

**1.3) Graph**

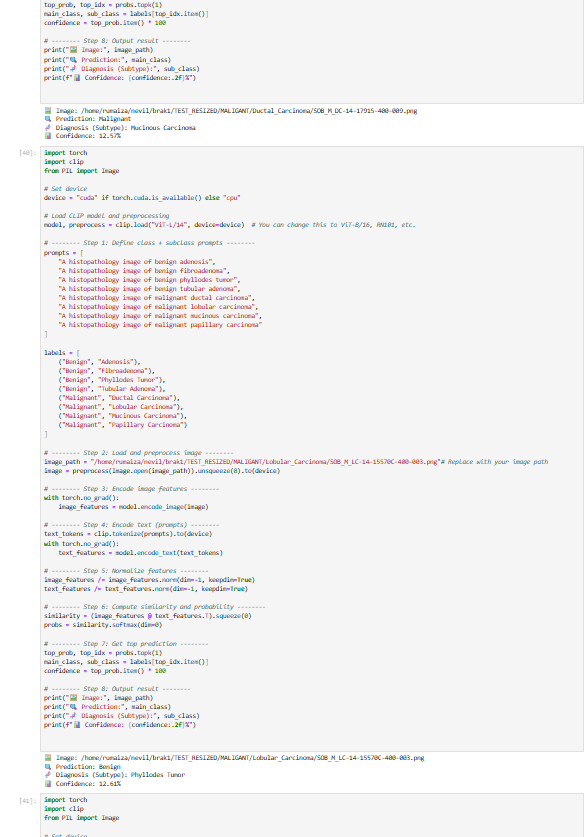


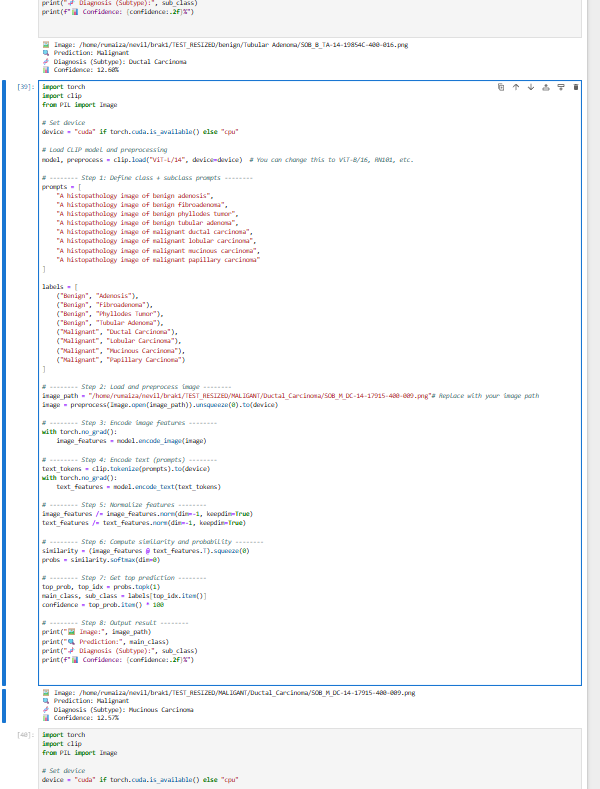
**7) Conclusion**

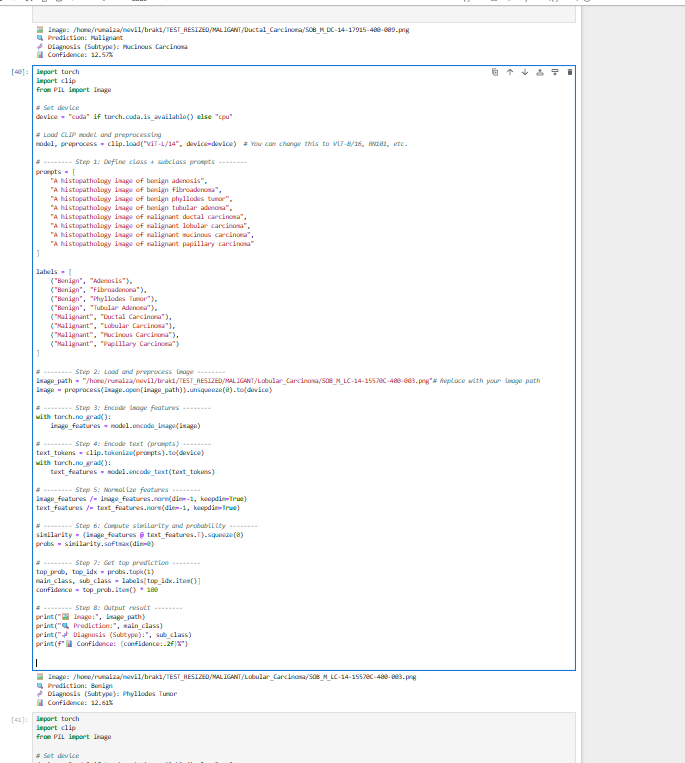
This project highlights the potential of vision–language models (VLMs) such as CLIP for histopathological image classification in breast cancer diagnosis. Both fine-tuned CLIP and ResNet50-based models demonstrated strong performance, achieving 88% to 92% accuracy across eight subtypes of benign and malignant breast lesions. Notably, CLIP in zero-shot mode generalized well when guided by clinically relevant text prompts, even without supervised training. These results affirm the power of language–image alignment in medical imaging tasks. Evaluation through confusion matrices and training-validation accuracy curves confirmed consistent learning and reliable generalization throughout the training process, even under varied data conditions.

During experimentation, it was observed that the model became overfitted when trained for too many epochs — achieving 100% training accuracy while validation accuracy plateaued at 89%, particularly after the 55th epoch. To address this, targeted data augmentation strategies, especially for the benign lesion classes, were employed. This approach helped mitigate overfitting and improved robustness, resulting in more balanced performance (training accuracy 99%, validation 89%). In inference, the model accurately predicted the main class (benign vs malignant) in most cases but showed occasional errors in subtype classification, highlighting the need for refined prompts and deeper architectural tuning.

Overall, this study establishes a strong foundation for integrating CLIP into clinical diagnostic workflows and opens avenues for future research in multi-modal, scalable, and privacy-preserving medical AI systems.







**8) Research paper analysis :**

**1. PathologyVLM (2025)**

* Abstract: Introduces a VLM tailored for pathology tasks like visual QA and classification.
* Methodology: Pretrained on image–text pairs using domain-aligned visual encoder and scale-invariant connector.
* Results: Outperforms CLIP/BioViL in zero-shot and few-shot classification across pathology benchmarks.

**2. Boosting VLMs for Histopathology (2023)**

* Abstract: Enhances CLIP/ALIGN zero-shot classification via patch-level inference.
* Methodology: Transductive learning using patch–patch and patch–text affinities at inference.
* Results: Accuracy improvements (≈10–15%) on BreakHis and other histology datasets.

**3. AI in Breast Cancer Pathology: Literature Review (2024)**

* Abstract: Reviews AI in pathology: image grading, mitotic detection, TILs.
* Methodology: PRISMA-based review of >70 studies on WSI.
* Results: Identifies interpretability and standardization as key gaps.

**4. Deep Learning in Breast Histopathological Imaging (2024)**

* Abstract: Surveys DL techniques (CNNs, Transformers) across diagnosis, grading, prognosis.
* Methodology: Organized by application: tumor detection, subtyping, etc.
* Results: ViT and ResNet dominate classification; attention needed for rare class detection.

**5. CBAM‑EfficientNetV2 for Histopathology (2024)**

* Abstract: Improves classification by combining CBAM attention with EfficientNetV2.
* Methodology: CLAHE for preprocessing, CBAM for attention refinement.
* Results: Achieved 99.1% accuracy, 98.3% F1-score on BreakHis dataset.

**6. SupCon-ViT: Supervised Vision Transformer (2024)**

* Abstract: Uses supervised contrastive learning on ViT for breast cancer images.
* Methodology: Trains ViT using SupCon loss; tested on IDC histopathology images.
* Results: F1-score 0.82, Specificity 0.90 with fewer annotations.

**7. Histopathological Diagnosis via Deep Mutual Learning (2024)**

* Abstract: Uses two CNNs that learn from each other to improve performance.
* Methodology: Joint optimization of twin models with cross-loss functions.
* Results: Outperformed baseline CNN by 3–5% on binary classification.

**8. Ensemble DL+ML on ResNet50 Features (2024)**

* Abstract: Combines ResNet50V2 features with ML models (SVM, XGBoost).
* Methodology: Uses deep CNN for feature extraction, ensemble ML for classification.
* Results: Imprxoved performance with AUC > 0.97 on BreakHis.

**9. DL for Breast Biopsy Risk Stratification (2024)**

* Abstract: Predicts high vs. low risk from preoperative biopsies.
* Methodology: DL model trained on annotated WSI with outcome labels.
* Results: AUC = 0.87 for invasive risk prediction.

**10. DL Risk Profiling vs Multigene Assay (2024)**

* Abstract: Compares DL histopathology-based risk prediction to gene assays.
* Methodology: Uses CNNs to infer recurrence risk.
* Results: DL shows 92% agreement with 21-gene assay.

**11. Predicting Gene Expression from Histology (2024)**

* Abstract: Predicts gene expression scores directly from WSIs.
* Methodology: CNNs trained to regress expression scores (e.g., proliferation index).
* Results: Pearson r > 0.7 for key genes.

**12. Federated Learning for TNBC (2023)**

* Abstract: Predicts treatment response in triple-negative breast cancer.
* Methodology: CNNs trained with federated learning across hospitals.
* Results: Maintained privacy while achieving >85% accuracy.

**13. MFF‑HistoNet with Quantum Tensor Modules (2025)**

* Abstract: Combines CNN with quantum-inspired tensors for robust classification.
* Methodology: Uses BiLSTM + QTM fusion for multi-class learning.
* Results: Achieved 97.6% classification accuracy.

**14. EXAONEPath: Patch-Level Foundation Model (2024)**

* Abstract: Stain-invariant model for general WSI understanding.
* Methodology: Self-supervised contrastive learning on pathology patches.
* Results: Strong generalization; >90% transfer accuracy on unseen organs.

**15. ViT–DeiT Ensemble for Histopathology (2022)**

* Abstract: Combines ViT and DeiT transformers for robust classification.
* Methodology: Ensemble via averaging predictions; no fine-tuning required.
* Results: 98.17% accuracy on multiclass breast histopathology images.

**16. Deep Learning for BC Detection with EfficientNet (2024)**

* Abstract: Uses pretrained EfficientNetB0 for binary cancer detection.
* Methodology: Image normalization, transfer learning, fine-tuning.
* Results: Achieved ≈94.5% accuracy.

**17. Nottingham Grade vs Gene Mutations (2023)**

* Abstract: Correlates histologic grade with genomic mutations like TP53, PIK3CA.
* Methodology: Links slide features with genomic panel data.
* Results: Moderate-to-high correlation supports multimodal diagnosis.

**18. Single-Cell & Spatial Transcriptomics (2024)**

* Abstract: Explores heterogeneity in breast cancer via scRNA and spatial mapping.
* Methodology: Tissue dissection → spatial barcoding → image fusion.
* Results: Reveals tumor microenvironment zones impacting prognosis.

**19. Explainable Mammogram Classifier (GMIC) (2020)**

* Abstract: Predicts malignancy while visualizing decisions.
* Methodology: CNN with heatmaps and clinical context input.
* Results: AUC 0.93; improved radiologist accuracy in experiments.

**20. CancerLLM / MedUnA (2024)**

* Abstract: Multimodal AI combining pathology images + reports + genes.
* Methodology: Uses large multimodal transformers + contrastive training.
* Results: 85–90% accuracy on multi-label breast cancer tasks.

**21. AI-based CEUS Analysis for Lymph Node Metastasis (2025)**

* Abstract: Combines CEUS images + clinical markers for ALN prediction.
* Methodology: CNN with handcrafted features (e.g., echo type).
* Results: 89.1% accuracy, outperforming radiologist baseline.

**22. Deep Features for 3D Histopathology Mapping (2023)**

* Abstract: Transforms 2D H&E into 3D volumetric reconstructions.
* Methodology: 3D CNN inference across z-stacks of histology images.
* Results: Enabled tumor volume estimation; useful for surgery planning.

**23. Marker Identification via DL (2023)**

* Abstract: DL used to find new IHC markers for early BC detection.
* Methodology: Trained CNNs on multiplexed IHC panels.
* Results: Identified markers with better early-stage sensitivity**.**

**24. DL + Graph Networks for Subtype Classification (2024)**

* Abstract: Combines CNNs with graph models for spatial cell interaction.
* Methodology: GraphSAGE used on cell graph from histology.
* Results: Boosted accuracy for Luminal vs. HER2 vs. Basal subtyping.

**25. AI to Match Histology and Radiomics (2024)**

* Abstract: Links WSI with MRI-based radiomics for whole-tumor insight.
* Methodology: Dual encoders with joint contrastive loss.
* Results: Helped bridge macro (MRI) and micro (histology) view of tumors.