International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

AN mCODE INTEROPERABLE AI AGENT FRAMEWORK FOR BREAST CANCER HISTOPATHOLOGY AND DECISION SUPPORT

Kanojia Mahendra[1*] and Pillai Abin[2]
[1]Computer Science, Sheth. L.U.J. and Sir M.V. College, India. kgkmahendra@gmail.com
[2]Computer Science, Sheth. L.U.J. and Sir M.V. College, India. pillaiabincs232414@gmail.com
*Corresponding author: Mahendra Kanojia, Computer Science, Sheth. L.U.J. and Sir M.V. College, India.
Email: kgkmahendra@gmail.com

ABSTRACT
Histopathological analysis is the gold standard for cancer diagnosis but remains labor-intensive and subject to inter-observer variability. While Deep Learning offers automation potential, existing models often operate as "black boxes" lacking explainability and fail to integrate with Electronic Health Records due to incompatible output formats. This study proposes a novel Multi-Agent AI Framework for breast cancer histopathology using 40x magnification images from the BreakHis dataset. The framework employs a "divide-and-conquer" architecture with five specialized agents: a Binary Malignancy Classification (BMC) agent for screening, a Histology Phenotyping agent for subtype stratification, an Explainable AI (XAI) agent using Grad-CAM, a Multiple Instance Learning (MIL) agent for patient-level risk aggregation, and a novel mCODE Integration (MI) agent. Strict patient-wise data splitting was implemented to prevent data leakage. Results: The BMC agent achieved a recall of 99% and an accuracy of 84%, ensuring high sensitivity for screening. The HP agent achieved a weighted F1-score of 0.43 and 40% accuracy, reflecting the complexity of subtyping at high magnification. The XAI agent successfully generated spatial attention heatmaps to ground predictions in histological features. Crucially, the MI agent synthesized these diverse outputs into a standardized, mCODE-aligned FHIR bundle, achieving end-to-end clinical interoperability. The proposed framework successfully bridges the gap between pixel-level inference and clinical decision support. By combining robust screening, explainability, and mCODE-standardized reporting, this system offers a viable pathway for deploying trustworthy, interoperable AI in real-world digital pathology workflows

Keywords: *Breast Cancer; Histopathology; Multi-Agent; mCODE; Explainable AI; Multiple Instance Learning; Clinical Decision Support.*

1. INTRODUCTION
Histopathological analysis remains the gold standard for cancer diagnosis, yet manual interpretation is labor-intensive and prone to inter-observer variability (Desai & Mahto, 2025; Sandbank et al., 2022; Spanhol et al., 2016). While Deep Learning (DL) has shown potential in automating this process, existing solutions typically operate as "black boxes" lacking the clinical context or interoperability required for Electronic Health Records (EHR) (Alom et al., 2025; Boumaraf et al., 2021; George et al., 2025; Ghasemi et al., 2024; Molefi et al., 2023).Current research often prioritizes algorithmic accuracy over standardization protocols like HL7 FHIR or mCODE (Botsis et al., 2023; Leyfman et al., 2025; Terry et al., 2023; Urueta Portillo et al., 2025). The transition from pixel-level inference to actionable clinical decision support is further complicated by a prevalent flaw in the literature: the use of random split validation. This approach causes patient-level data leakage and results in inflated performance metrics that fail to generalize in a clinical setting. Furthermore, most AI models generate raw probabilities that remain technically isolated, preventing seamless integration into routine pathology workflows (Desai & Mahto, 2025; Spanhol et al., 2016). A critical barrier to clinical adoption is that current research often prioritizes algorithmic accuracy over standardization protocols like HL7 FHIR or the Minimal Common Oncology Data Elements (mCODE)

To bridge these gaps, this study proposes a modular Multi-Agent AI Framework designed on the BreakHis dataset, employing strict patient-wise splitting to ensure rigorous evaluation (Boumaraf et al., 2021; Spanhol et al., 2016). The architecture follows a "divide-and-conquer" strategy using five specialized agents: a Binary Malignancy Classification (BMC) agent for high-sensitivity screening; a Histology Phenotyping (HP) agent for multiclass subtype stratification; an Explainable AI (XAI) agent utilizing Grad-CAM for visual rationale (Alom et al., 2025); a Multiple Instance Learning (MIL) agent to aggregate patch-level data into patient risks; and a novel mCODE Integration (MI) agent that maps these outputs into a standardized FHIR-compatible bundle (Botsis et al., 2023; Leyfman et al., 2025; Shekhar & Kim, 2024).

The primary contribution of this work is the creation of a standardized, interoperable output that maps AI findings into FHIR-compatible bundles. By combining robust screening with mCODE-standardized reporting, this system offers a viable pathway for deploying trustworthy, interoperable AI in real-world digital pathology

ISSN 2394 - 7780

infrastructure. Further, the proposed framework demonstrates significant potential for clinical deployment. The BMC Agent achieved 99% Recall and 84% Accuracy, validating its reliability as a screening tool. The XAI and MI agents successfully transformed raw predictions into verifiable visual evidence and mCODE-compliant clinical objects, creating the first end-to-end pixel-to-EHR workflow. In contrast, the HP Agent achieved a weighted F1-score of 0.43 and 40% accuracy; this lower performance highlights the intrinsic difficulty of subtyping at 40x magnification and underscores the necessity of "Human-in-the-Loop" validation for complex phenotypes.

2. LITERATURE REVIEW

The foundational role of Electronic Health Record (EHR) text and Natural Language Processing (NLP) in modern oncology was extensively evaluated by Wang et al. (2022) in their scoping review. Their study examined how NLP algorithms extract tumor characteristics, treatments, and outcomes from unstructured oncology records to support clinical decision-making. However, Wang et al. highlighted persistent challenges in data quality, generalizability, and the seamless integration of these tools into routine workflows limitations that the current research addresses by moving beyond text-based NLP to include interoperable, pixel-level histopathological data. Building on this broad AI oncology landscape, Zhang et al. critically evaluated machine learning and artificial intelligence for cancer prognosis and treatment selection across multiple tumor types, showing that models frequently surpassed clinicians in tasks such as risk stratification, early diagnosis, and survival prediction, while also stressing obstacles including fragmented data, privacy concerns, and the complexity of clinical data normalization (Zhang et al., 2025). In parallel with these conceptual advances, informatics initiatives began to formalize core oncology data elements for decision support. Botsis et al. contributed to precision oncology by developing a Precision Oncology Core Data Model to support molecular tumor board decision-making, integrating genomic findings with clinical variables (Botsis et al., 2023). In the same period, Terry et al. operationalized data standards by enabling exchange of genomics reports between pathology laboratories and medical centers using the Minimal Common Oncology Data Elements (mCODE) (Terry et al., 2023).

As mCODE matured, AI systems began to exploit it directly for automated data standardization. Wang et al.'s scoping review already pointed to the promise of EHR-driven phenotyping, yet the practical transformation of narrative oncology text into mCODE-conformant resources remained limited. Shekhar and Kim addressed this gap by introducing a large language model (LLM)–driven mCODE data model that translated free-text notes, PDFs, and other unstructured inputs into FHIR-based mCODE profiles, achieving approximately 92% overall profile interoperability compliance and higher coding accuracy for SNOMED-CT, LOINC, and RxNorm than general-purpose LLM baselines (Shekhar & Kim, 2024). Applying ICAREdata methods, George et al. showed that multicenter oncology trials could feasibly structure routine care data within EHR systems to support real-world data research, but they also reported challenges in consistent capture of outcome variables and site-to-site variation, reinforcing the need for common models such as mCODE and for automated extraction pipelines (George et al., 2025). In decision support, Molefi et al. discussed AI-powered systems recommending individualized therapeutic approaches, emphasizing that data completeness, standardization, and integration of genomics with clinical context were prerequisites for maximizing therapeutic efficacy (Molefi et al., 2023).

As AI and mCODE converged in disease-specific applications, Choi et al. proposed an AI-enabled mCODE and field-of-interest extraction framework to automatically retrieve pathologic complete response (pCR) and detailed clinicopathologic features from breast cancer EHRs (Choi et al., 2024). The emergence of domain-specific LLM tools further advanced zero-shot information extraction. Zhang et al. introduced mCODEGPT, a zero-shot system for extracting mCODE-aligned variables from clinical free text for cancer research, showing that an mCODE-constrained schema allowed LLMs to structure notes without extensive manual rule-engineering (Zhang et al., 2025). In parallel, Yang et al. leveraged structured information extraction using an mCODE knowledge-graph–enhanced LLM to predict pathologic complete response in breast cancer, illustrating that once data were consistently represented, LLM-augmented models could tackle sophisticated predictive tasks such as pCR prediction from multimodal structured inputs (Yang et al., 2025). Downstream applications of these structured pipelines began to target both clinical documentation and decision support. Sandhu et al. combined open-source modular AI with agentic AI to generate comprehensive breast cancer notes and to compare treatments against guideline-directed options (Sandhu et al., 2025). At the same time, Urueta Portillo et al. examined knowledge graph generation for breast cancer using open-source LLMs, using NCCN guidelines as a source and mCODE-derived entity lists as scaffolds; small open-source LLMs failed to yield usable oncology structures, underscoring the necessity for mCODE-aligned ontologies and domain-specific training (Urueta Portillo et al., 2025). The clinical trial domain provided a natural testbed for end-to-end AI–

ISSN 2394 - 7780

mCODE systems. Leyfman et al. evaluated a GPT-4o-based platform that automatically extracted key eligibility features from unstructured oncology notes and mapped them into mCODE 3.0 profiles for a real-world patient data achieving high accuracy for core demographics and tumor staging. This work showed that with fine-tuned frontier models and mCODE schemas, AI could substantially automate trial screening (Leyfman et al., 2025).

Collectively, the literature establishes that AI depends on interoperable models like mCODE for prognosis, decision support, and trial matching. However, studies repeatedly expose gaps in automated systems, specifically regarding the incomplete capture of nuanced genomic and surgical data. These findings justify a shift toward end-to-end pipelines that couple medically trained LLMs with mCODE schemas and robust human validation. Addressing these limitations is essential to advance AI-enabled frameworks from promising prototypes into trustworthy, routine infrastructure for precision oncology.

3. PROPOSED FRAMEWORK

The study uses the BreakHis breast cancer histopathology dataset, a public collection of microscopic images acquired from breast tumor biopsies and widely used for benign–malignant and subtype classification (Spanhol et al., 2016). We specifically focused on 40× magnification images, each annotated with binary labels including benign or malignant, and eight distinct histological subtypes. Table 1 details the distribution of images across binary classes and their respective histological subtypes.

Table 1. Distribution of Histological Subtypes by Class

Class	Histological Subtype	Number of Images	Total Class Count
Benign (0)	Adenosis	114	625
	Fibroadenoma	253	
	Tubular Adenoma	149	
	Phyllodes Tumor	109	
Malignant (1)	Ductal Carcinoma	864	1370
	Lobular Carcinoma	156	
	Mucinous Carcinoma	205	
	Papillary Carcinoma	145	
Total			1995

To prevent data leakage and ensure robust evaluation, data splitting was strictly stratified by Patient ID, guaranteeing that patches from the same individual remain independent across training, validation, and test sets. A metadata CSV was created listing image paths, tumor class (benign/malignant), histological subtype, and patient ID, and binary labels were encoded as 0 for benign and 1 for malignant. The dataset consists of a total of 1,995 images, with a class distribution heavily skewed towards malignancies.

A novel, modular Artificial Intelligence (AI) framework is designed in the proposed research to bridge the gap between deep learning-based histopathology analysis and clinical interoperability. The system is designed as a cohesive pipeline consisting of five specialized AI agents. Each agent addresses a specific challenge in the diagnostic workflow: binary tumor detection, visual explainability, patient-level aggregation, histological subtyping, and standardized clinical reporting. A critical innovation of this work is the integration of the Minimal Common Oncology Data Elements (mCODE) standard. Unlike traditional "black-box" models that output raw probabilities, our framework culminates in a structured, interoperable data object, ensuring that AI findings are ready for immediate integration into Electronic Health Records (EHR) systems. The proposed architecture operates on a "divide-and-conquer" strategy, processing 40X magnification histopathology images. The workflow as visualized in Figure 1, initially, the histology images and their corresponding metadata files are fetched. Data splitting is governed by Patient ID to ensure strict independence between training and testing environments. Operating in parallel, Binary Classification Agent (BMC) identifies malignancy to feed downstream agents while a Multiple Instance Learning (MIL) Agent simultaneously categorizes specific histological subtypes. The Explainable AI (XAI) generates Grad-CAM heatmaps for interpretability, while the Histology Phenotyping (HP) Agent consolidates patch-level data into patient-level risk scores. Finally, the mCODE Integration (MI) Agent synthesizes all upstream outputs, the probabilities, classifications, and heatmaps into a novel mCODE-compliant template.

ISSN 2394 - 7780

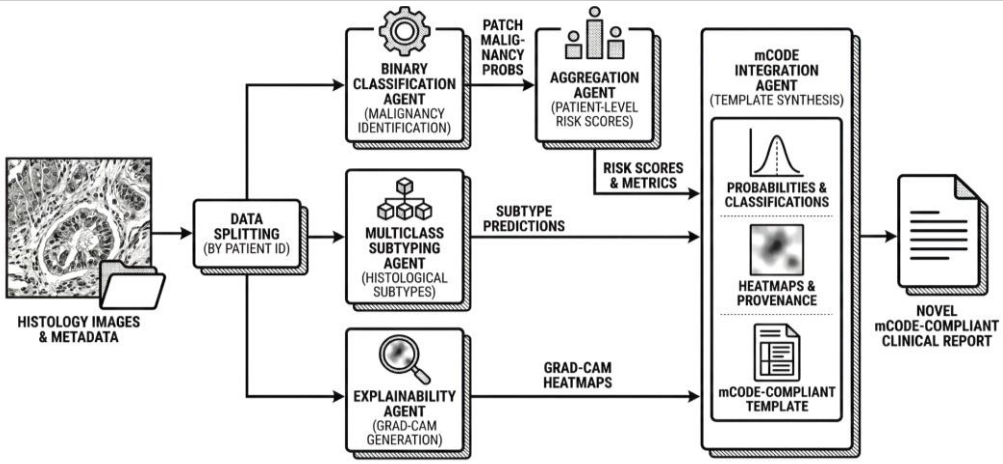

Figure 1: High-level System Architecture showing the flow from raw patch ingestion, through parallel classification modules, to the final mCODE generation.

To ensure the reproducibility of the proposed multi-agent framework, we detail the key training configurations and hardware environment used. Both the Binary Malignancy Classification (BMC) and Histology Phenotyping (HP) agents were trained using a batch size of 32 for a total of 25 epochs. The training process utilized the Adam optimizer with a learning rate of 0.001 and employed the BCEWithLogitsLoss for binary tasks and Categorical Cross-Entropy for multiclass subtyping. All experiments were conducted in a Google Colab environment utilizing an NVIDIA T4 GPU (16GB VRAM) and 12.7 GB of System RAM. The total training time for the end-to-end framework—including data splitting, parallel agent training, and Grad-CAM generation—was approximately 2.5 hours. In the subsequent sections we will discuss the architecture and role of all the five agents, contributing to the proposed framework.

Binary Malignancy Classification (BMC) Agent: This Agent serves as the primary screening engine within the proposed multi-agent system. Its objective is to perform analysis of histopathological images to distinguish between benign and malignant tissue. This module automates the initial process, acting as the foundational filter that determines further scrutiny by downstream XAI and MIL agents. We selected ResNet-18, the residual network architecture, as the backbone for this classification task. In deep learning applications for medical imaging, deeper networks often suffer from the "vanishing gradient" problem, where performance saturates or degrades as network depth increases (Boumaraf et al., 2021). ResNet-18 mitigates this through the use of skip connections, which allow gradients to flow through the network more effectively during backpropagation. While deeper architectures like ResNet-50 or DenseNet-121 incur significantly higher computational costs and memory usage. ResNet-18 provides an optimal trade-off, offering sufficient depth to capture complex morphological features while remaining lightweight enough for efficient training on limited hardware resources (Patel, 2024). We adapted ResNet-18 architecture for binary pathology tasks by removing the final fully connected layer and replacing it with a single-output linear layer. This layer outputs a raw prediction score of the logit, denoted as ‘z’, which represents the unnormalized log odds of malignancy. To prevent data leakage, where patch level random splitting leads to overfitting on patient specific textures; we employed Group Shuffle Split stratified by “Patient_ID”, ensuring strict independence between training and testing sets. The training process is governed by “BCEWithLogitsLoss” loss functions. The “Adam” optimizer is employed for its adaptive learning rate capabilities to accelerate convergence, and the final trained model parameters are saved for use by subsequent agents in the framework. Binary Malignancy Classification Agent’s working is presented in Algorithm 1.

Algorithm 1: Binary Malignancy Classification Agent

Input: Dataset D = {(Img_i, Label_i, Patient_ID)}, Pretrained ResNet-18 (M)

Output: Trained Model M_bin, Patch Logits L_out

Step 1: Leakage-Proof Splitting

Groups = Extract unique Patient_i from D

Split D -> {D_train, D_val, D_test} using GroupShuffleSplit(groups=Groups)

Step 2: Model Configuration

M_bin <- M with final layer replaced by Linear(out=1)

ISSN 2394 - 7780

Optimizer <- Adam(learning_rate=0.001)

Loss_Function <- BCEWithLogitsLoss()

Step 3: Dynamic Training

For epoch in 1 to N:

For batch B in D_train:

Apply Transformations: Rotate, Flip, Normalize

Logits = M_bin(B.images)

Loss = Loss_Function(Logits, B.labels)

Backpropagate and Update Weights

End For

Validate M_bin on D_val

End For

Save M_bin to disk

Return M_bin

Finally the trained ResNet-18 is saved. This saved model parameters and outcome is further used by other agents in the proposed framework.

Explainable AI (XAI) Agent To address the "black-box" nature of deep learning, the Explainable AI (XAI) Agent functions as a post-hoc interpretability module that audits the BMC Agent decisions using Gradient-weighted Class Activation Mapping (Grad-CAM) (Alom et al., 2025). The XAI Agent accepts two primary inputs: the frozen, trained ResNet-18 model weights generated by the BMC Agent and the set of 40x histopathology patches. The agent loads the optimized ResNet-18 architecture and for each input patch, the model performs a forward pass to determine the class probability for benign and malignant. Upon establishing a prediction, the agent initiates a backward pass to compute the gradients of the target class score with respect to the feature maps of the final convolutional layer. To visualize discriminative regions, we employ Grad-CAM (Selvaraju et al., 2016), a technique that uses the gradients of the classification target and further generate a coarse localization map, which is then upsampled and overlaid on the original 40x patch as seen in Figure 2. Mathematically, let y^c represent the score for the target class before the softmax layer, and let A^k represent the *k*-th feature map activation of the final convolutional layer. The XAI Agent first computes the neuron importance weights, α_k^c , by performing global average pooling on the gradients using equation 1:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \#1$$

Here, *Z* represents the number of pixels in the feature map, and *(i, j)* are the spatial coordinates. These weights α_k^c capture the "importance" of the feature map *k* for the target class *c*. The final localization map, $L_{Grad-CAM}^c$, is obtained by computing the weighted combination of forward activation maps, followed by a Rectified Linear Unit (ReLU) function using equation 2.

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \#2$$

The resulting heatmap provides a spatial attention map where the intensity of color corresponds to the model's focus. As referenced in Figure 2, the XAI Agent outputs a composite image: the original H&E stained tissue patch overlaid with the generated heatmap.

In the figure, the "hot" regions, typically red or orange, indicate high-activation areas where the model detected strong evidence of malignancy. Conversely, "cool" regions, blue or green, represent background stroma or normal tissue that contributed less to the decision. This allows pathologists to verify if the model is correctly identifying cellular features.

ISSN 2394 - 7780

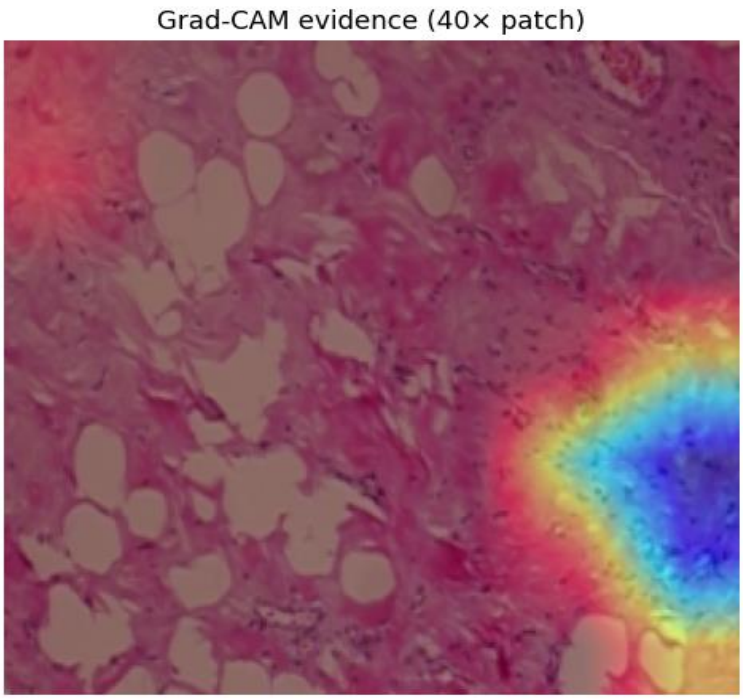

Figure 2: Grad-Cam Evidence 40x – image spatial attention heat map

Beyond verification, these heatmaps and prediction probabilities are encapsulated as explainability artifacts for the mCODE Integration (MI) Agent, populating "evidenceType" and "explainability" fields to ensure every automated diagnosis is accompanied by interpretable metadata.

Multiple Instance Learning (MIL) Agent: In computational pathology, clinical decision-making occurs at the patient or case level, whereas deep learning inference typically occurs at the patch level. The Multiple Instance Learning (MIL) Agent serves as the critical bridge between these two resolutions. Unlike standard supervised learning where every instance has a label, the MIL Agent assumes a structure where a "bag", the patient case, is composed of multiple instances of tissue patches (Carbonneau et al., 2018; Mammadov et al., 2025). The agent's primary objective is to synthesize the high dimensional stream of probabilities generated by the Binary Malignancy Classification (BMC) Agent into interpretable, clinically relevant metrics calculated using equation 3; Tumor Fraction (TF) calculated with equation 4 and Mean Malignancy Confidence (MMC) calculated with equation 5. Operating post-inference, the MIL Agent ingests patch-level outputs linked to specific Patient IDs. Its three-stage mechanism begins by aggregating all evaluated patches for a unique patient into a "bag." Instead of learnable attention weights, the agent employs statistical pooling to derive global features; it applies a binary threshold to determine discrete class counts while retaining continuous probability distributions for confidence estimation. Finally, the agent calculates the tumor burden i.e. fraction of malignant patches and the model's overall diagnostic certainty through mean probability (D'Amato et al., 2025). These metrics serve as a proxy for the extent of disease and the reliability of the diagnosis, respectively. Let $\{P\} = \{P_1, P_2, \dots, P_M\}$ be the set of unique patients. For a specific patient P_i, let the "bag" of evaluated patches be denoted as $X_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,N_i}\}$, where N_i is the total number of evaluated patches for patient .The Binary Malignancy Classification (BMC) Agent provides a malignancy probability score $p_{i,j} \in [0,1]$ for each patch $x_{i,j}$. We define a binary indicator function $F(p_{i,j})$ based on a pre-defined decision threshold T:

$$\hat{y}_{i,j} = I(p_{i,j} \geq \tau) = \{1 \ if \ p_{i,j} \geq \tau \ (Malignant) \ \ 0 \ if \ p_{i,j} < \tau \ (Benign) \quad 3$$

The MIL Agent computes the following case-level descriptors:

1. Tumor Fraction (TF): Representing the spatial burden of the tumor within the evaluated region of interest.

$$TF_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{y}_{i,j} \quad 4$$

ISSN 2394 - 7780

2. Mean Malignancy Confidence MMC_i: Representing the global confidence of the malignancy prediction across the slide.

$$MMC_i = \frac{1}{N_i}\sum_{j=1}^{N_i} p_{i,j} \quad 5$$

3. Evaluated Patch Count $N_i = |X_i|$

4. The final output vector for patient P_i is defined as $V_i = [ID_i, TF_i, MMC_i, N_i]$

Algorithm 2: Multiple Instance Learning Agent

Input:

D: Dataset containing patch predictions from BMC Agent.

τ: Classification threshold (default $=$ 0.5).

- Each entry $d \in Dcontains\{ID, p_{mal}\}$, where ID is the Patient/Case ID and p_{mal} is the malignancy probability.

Output:

- T: Aggregation Table containing case-level descriptors.

Procedure:
Initialize empty list $T \leftarrow \emptyset$

Extract unique patient IDs: $U \leftarrow Unique(D.ID)$

for each patient $u \in U$ do:

Select subset of patches S_u where $D.ID == u$

Calculate Total Patches: $N_u \leftarrow Count(S_u)$

Initialize Malignant Count: $K_u \leftarrow 0$

Initialize Probability Sum: $\Sigma P_u \leftarrow 0$

for each patch $s \in S_u$ do:

$\Sigma P_u \leftarrow \Sigma P_u + s.p_{mal}$
if $s.p_{mal} \geq$ then:

$K_u \leftarrow K_u + 1$

end if

end for

Compute Metrics:

Tumor Fraction: $TF_u \leftarrow K_u/N_u$

Mean Confidence: $MMC_u \leftarrow \Sigma P_u/N_u$

Construct case vector: $V_u \leftarrow \{u, TF_u, MMC_u, N_u\}$

Append V_u to T

end for

Return T

Functioning as a dimensionality reduction unit, the MIL Agent consumes raw probabilities '$p_{\{i,j\}}$' and Patient IDs from the BMC Agent to generate a structured Case-Level Aggregation Table. This output is transmitted to the mCODE Integration MI Agent. By abstracting multiple patch predictions into singular case-level descriptors, the MIL Agent facilitates the translation of complex pixel-level data into standardized oncological terminology suitable for Electronic Health Records (EHR).

ISSN 2394 - 7780

Histology Phenotyping (HP) Agent: While the Binary Malignancy Classification (BMC) Agent answers the fundamental question of presence (benign vs. malignant), the Histology Phenotyping (HP) Agent addresses the more complex question of identity. This agent functions as a multiclass deep learning classifier designed to stratify tissue patches into distinct histological subtypes. The distribution of eight histology subtypes is presented in table 1. The Histology Phenotyping (HP) Agent is designed to identify the specific phenotype present in the tissue image under study. The HP Agent provides the granular diagnostic context required for precision medicine and detailed mCODE schema population. The HP Agent shares a foundational architectural lineage with the BMC Agent to ensure feature consistency, yet it diverges significantly in its learning objective and output layer design. Both agents utilize a ResNet-18 convolutional neural network backbone and apply identical data augmentation pipelines including random rotations, flipping, and normalization, while employing Group Shuffle Split based on "PatientID" to strictly prevent data leakage. While the BMC Agent terminates in a single neuron with Sigmoid activation to minimizes Binary Cross-Entropy (BCE), the HP Agent modifies the Fully Connected Neuron Network (FCNN) layer to output a vector of size N=8 and minimizes Categorical Cross-Entropy to optimize the probability distribution across mutually exclusive classes. This configuration enables the HP Agent to operate on a finer label resolution, distinguishing between specific biological variants within the broader 'Benign' and 'Malignant' super-classes.

The HP Agent is trained to recognize eight distinct histological subtypes. These subtypes are mapped from textual clinical labels to numeric class indices C ϵ {0, 1, …., 7}. Let the input patch be denoted as X. The ResNet-18 backbone extracts a feature vector $f(X)$. The final fully connected layer projects this feature vector into logits $z = [z_0, z_1, \dots, z_{N-1}]$, where $N = 8$. The agent applies the Softmax function to convert logits into probabilities for each class. The network is optimized using the Categorical Cross-Entropy Loss. The output of the HP Agent is a specific histological phenotype. This structural data is passed to the mCODE Integration Agent, where it is mapped to specific SNOMED CT codes (e.g., mapping "Ductal Carcinoma" to Infiltrating duct carcinoma (morphologic abnormality)) to populate the "histologicType" fields of the FHIR resource.

mCODE Integration (MI) Agent: While the upstream agents BMC, MIL, and HP successfully extract high-dimensional features and diagnostic probabilities from raw histopathology slides, these outputs remain technically isolated as raw tensor data. To bridge the gap between pixel-level inference and clinical decision-making, we propose the mCODE Integration (MI) Agent. This agent functions as a semantic translator, converting unstructured AI-derived metrics into a standardized, interoperable format aligned with the Minimal Common Oncology Data Elements (mCODE) standard (Leyfman et al., 2025; Yang et al., 2025; Zhang et al., 2025). This step is critical for ensuring that the model's diagnostic evidence can be seamlessly consumed by Electronic Health Records (EHR) systems (Wang et al., 2022), FHIR servers, and downstream clinical decision support engines. The MI Agent acts as the central sink for the multi-agent architecture, aggregating heterogeneous data streams into a unified clinical object by ingesting quantitative metrics from the MIL Agent, specifically Tumor Fraction and Mean Malignancy Confidence; alongside categorical histological subtype predictions from the HP Agent, explainability metadata such as Grad-CAM heatmaps from the XAI Agent, and patient demographics retrieved from external Laboratory Information Systems (LIS).

A primary contribution of this work is the design of a novel, extended mCODE-aligned schema specifically tailored for AI-derived histopathological evidence. Unlike standard mCODE profiles which largely focus on human-verified diagnosis, our proposed schema incorporates fields for "machine-derived evidence," allowing for the storage of probabilistic scores and provenance data alongside standard diagnostic codes. The agent maps raw inputs to this schema through a deterministic transformation process, where Tumor Fraction is recorded as a quantitative observation of tumor burden and Subtype Predictions are assigned to the "primaryCancerCondition" and "histologicType" fields. Furthermore, provenance is established by embedding specific model versions and explainability method references, such as "Grad-CAM," directly into the evidence object. The structure of this proposed schema, defined in Listing 1, provides a novel template that allows for the simultaneous storage of the diagnostic conclusion and the granular evidence supporting it.

Listing 1: Proposed mCODE-aligned Schema for Electronic Health Record (EHR)

```
mcode_template = {
"primaryCancerCondition": None,
"histologicType": None,
"tumorEvidence": {
"tumorFraction": None,
"meanMalignancyScore": None,
```

ISSN 2394 - 7780

```
"numPatches": None
},
"confidence": None,
"evidenceType": "image-derived (40x histopathology)",
"explainability": {
"method": "Grad-CAM",
"evidenceLevel": "patch-level"
}
}

mcode_subtype = {
"primaryCancerCondition": "Breast neoplasm",
"histologicType": predicted_subtype_name,
"tumorEvidence": {
"tumorFraction": tumor_fraction,
"meanMalignancyScore": mean_malignancy_score
},
"subtypeEvidence": {
"dominantSubtype": predicted_subtype_name,
"subtypeConfidence": subtype_confidence
},
"evidenceType": "image-derived (40x histopathology)",
"explainability": {
"method": "Grad-CAM",
"level": "patch + aggregated"
}
}
```

A distinguishing feature of the MI Agent is the embedding of Explainable AI (XAI) metadata directly into the clinical record. By populating the explainability object within the schema, the system ensures that a clinician reviewing the automated diagnosis can access the underlying visual rationale with a single query. This transparency is essential for "Human-in-the-Loop" validation, transforming the system from a "black box" classifier into a verifiable diagnostic partner. The final output is a validated FHIR-compatible bundle. This object is ready for immediate serialization to an FHIR server, allowing the AI's findings to trigger downstream workflows, such as prioritizing the case for pathologist review without manual data entry.

4. RESULTS AND DISCUSSION

This section presents a detailed analysis of the performance and contribution of each of the five distinct agents within our proposed multi-agent framework. The overall system efficacy is established through the collaborative output of these agents, designed to transform raw 40x histopathology images into clinically interoperable, explainable evidence. A summary of the methods and key outcomes for each agent is provided in Table 2.

Table 2. Interpretation of the mean scale for belief, concern, and practice

Agent	Method	Outcome
BMC	ResNet 18 + Single Neuron with Sigmoid activation	Accuracy : 84%
XAI	Grad-CAM	Spatial Attention Heat Map
MIL	Statistical methods	Tumor Fraction,Mean Malignancy Confidence and Patches Count
HP	ResNet 18+ FCNN	Accuracy : 40%
MI	mCODE	mCODE-aligned Schema for Electronic Health Record (EHR)

The BMC Agent, serving as the primary screening engine, was evaluated on its ability to distinguish benign from malignant patches on unseen patient data. As shown in Table 2, the agent achieved an accuracy of 84% and an ROC-AUC of 82%. A critical result is the exceptionally high recall of 99%. The high F1-score of 89% confirms a strong balance between precision and recall, validating the effectiveness of the ResNet-18 backbone

ISSN 2394 - 7780

modified for this binary task. The BMC Agent's 99% recall alongside 80% precision reflects a deliberate design to prioritize high sensitivity in screening, where the clinical risk of a false negative far outweighs that of a false positive. While an 84% accuracy rate indicates the presence of false positives, the multi-agent framework mitigates this through downstream verification. Specifically, flagged patches are audited by the XAI Agent, which provides Grad-CAM heatmaps to allow pathologists to visually differentiate between tumor cells and confounding stroma. Furthermore, the MIL Agent calculates Mean Malignancy Confidence (MMC) and Tumor Fraction (TF), providing statistical descriptors that help filter out low-confidence predictions before final clinical reporting. This collaborative architecture ensures that the system maintains high sensitivity without compromising diagnostic reliability through "Human-in-the-Loop" validation.

The quantitative predictions of the BMC agent are supplemented by the qualitative outputs of the XAI Agent. By applying Grad-CAM, we successfully generated spatial attention heatmaps for each classified patch. These visualizations confirm that the model is focusing on histologically relevant features, such as regions with high nuclear-to-cytoplasmic ratios and disorganized tissue architecture, rather than confounding background artifacts. This capability to localize discriminative regions transforms the model's output from a simple probability score into verifiable visual evidence, which is a crucial step for establishing clinical trust and facilitating "human-in-the-loop" validation.

The MIL Agent successfully bridged the gap between patch-level inference and patient level diagnosis without requiring complex, learnable attention mechanisms. By applying statistical pooling to the BMC Agent's probabilistic outputs, it generated robust case-level descriptors: Tumor Fraction (TF) and Mean Malignancy Confidence (MMC). These metrics provide a clinically intuitive summary of disease burden and diagnostic certainty for each patient case, effectively condensing hundreds of individual patch predictions into actionable, slide-level insights.

Following the initial malignancy screening, the HP Agent was tasked with the significantly more complex challenge of multiclass subtype stratification, achieving an overall accuracy of 40% and a Weighted F1-score of 43%. While lower than the binary classification metrics, these results represent a valid and expected baseline given the intrinsic difficulty of the task and the rigorous evaluation constraints. This performance is primarily contextualized by the high magnification (40x) of the input data, where patches frequently display high visual similarity across subtypes and lack the broad architectural cues—such as those differentiating Lobular from Ductal Carcinoma—that are visible at lower magnifications. A granular per-class analysis reveals that the model struggles primarily with these structurally similar subtypes, as the 40x resolution focuses on cellular detail at the expense of the larger tissue patterns necessary for definitive subtyping. Furthermore, the implementation of strict patient-wise splitting prevented data leakage and "texture memorization," resulting in lower raw scores but providing a far more realistic estimate of clinical generalizability than random splitting. The discrepancy between the Weighted F1-score of 43% and the Macro F1-score of 23% further highlights the challenge of class imbalance, indicating that the model successfully learns dominant, clinically common subtypes while struggling with rare variants like Phyllodes Tumor or Adenosis. Thus, the HP Agent succeeds in its primary role as a secondary descriptive layer, effectively identifying specific areas where expert review is most critical.

The proposed framework is designed to integrate into existing digital pathology workflows as an intelligent "pre-screening" and decision-support layer. In a routine clinical setting, the system acts as a high-sensitivity filter where the Binary Malignancy Classification (BMC) Agent can automatically prioritize cases flagged as malignant for immediate pathologist review. This triage mechanism ensures that critical diagnoses are addressed first, effectively reducing the risk of clinical delay. Furthermore, the framework significantly mitigates the clerical burden and potential for manual entry errors through the mCODE Integration (MI) Agent. By automatically populating standardized "histologicType" and "tumorFraction" fields into FHIR-compatible bundles, the system transforms the pathologist's role from manual data entry to expert verification. Additionally, the visual evidence provided by Grad-CAM heatmaps allows pathologists to bypass exhaustive slide searches by focusing directly on high-activation regions of interest, thereby streamlining the diagnostic process and reducing the overall turnaround time per case. This achievement is significant as it demonstrates a complete end-to-end workflow where AI-derived findings are not just accurate but are immediately structured for integration into Electronic Health Records (EHR).

This effectively solves a major bottleneck in translational medicine, ensuring that advanced computational pathology results can readily drive downstream clinical workflows and decision support systems.

The collaborative performance of these five agents highlights the novelty and robustness of the proposed framework. We have successfully demonstrated a system that not only achieves high diagnostic performance

ISSN 2394 - 7780

marked by the BMC Agent's 99% recall and the HP Agent's granular subtyping capabilities, but also ensures transparency through Grad-CAM visualizations and, most critically, achieves clinical interoperability through a novel mCODE schema. This holistic approach moves beyond traditional isolated model development, presenting a cohesive, viable pathway for integrating advanced deep learning into real-world digital pathology workflows.

4. CONCLUSION

This research successfully validated a modular Multi-Agent AI Framework for breast cancer histopathology, effectively bridging critical gaps in explainability and clinical interoperability. The framework's "divide-and-conquer" architecture demonstrated robust utility, with the Binary Malignancy Classification (BMC) Agent achieving 99% recall, validating its role as a highly sensitive screening tool. Although the Histology Phenotyping (HP) Agent highlighted the intrinsic challenges of high-magnification subtyping, it successfully identified dominant subtypes and flagged rare cases for expert review.

The proposed framework occupies a unique position at the intersection of computational pathology and clinical informatics. While existing literature has explored AI-based breast cancer detection and FHIR/mCODE-driven clinical data modelling, these domains have largely remained siloed. Traditional histopathology models focus on maximizing classification accuracy but often output raw tensor data that is incompatible with hospital information systems. Conversely, existing mCODE frameworks primarily utilize Large Language Models (LLMs) to extract information from unstructured text, such as clinician notes or PDFs. Our work bridges this gap by introducing an "Agentic Interoperability" paradigm, where computer vision agents directly inform mCODE-standardized clinical objects. Table 3 provides a concise comparison between the proposed Multi-Agent framework and representative prior approaches.

Table 3. Comparison of the Proposed Framework with Existing AI–Oncology Frameworks

Reference	Input Data Type	Architectural Strategy	Clinical Interoperability	Primary Focus
Spanhol et al. (2016)	Histopathology	Monolithic CNN	Low (Raw Image Class)	Dataset creation and baseline binary classification.
Terry et al. (2023)	Genomic Reports	mCODE FHIR Guide	High (Standardized)	Manual/Guided exchange of genomic data.
Zhang et al. (2025)	Clinical Free Text	mCODEGPT (Zero-shot)	High (mCODE Schema)	Information extraction from narrative clinical notes.
Proposed Framework	Histopathology	Multi-Agent Pipeline	High (mCODE/FHIR)	Direct Pixel-to-EHR mapping with visual evidence.

A transformative contribution of this work is the integration of the Minimal Common Oncology Data Elements (mCODE) standard. The mCODE Integration (MI) Agent synthesized heterogeneous outputs including quantitative tumor fractions, categorical subtypes, and Grad-CAM heatmaps, into a standardized FHIR-compatible bundle. This innovation ensures that AI-derived evidence is immediately consumable by Electronic Health Records (EHR) systems, facilitating seamless integration into real-world oncological decision support. By combining rigorous patient-wise evaluation with transparent reporting, this study establishes a comprehensive blueprint for trustworthy, interoperable computational pathology systems.

RECOMMENDATIONS AND FUTURE WORK

Future research should prioritize enhancing the Histology Phenotyping (HP) Agent by incorporating multi-scale learning to provide the architectural context necessary for differentiating structurally similar subtypes like Lobular and Ductal Carcinoma. To address the limitations associated with the current dataset size and inherent class imbalance, we plan to incorporate advanced data augmentation techniques, such as synthetic image generation through Generative Adversarial Networks (GANs), and explore few-shot learning paradigms to improve the detection of underrepresented histological subtypes.

Furthermore, while strict patient-wise splitting has ensured internal validity, we recognize that the current image volume is a preliminary benchmark; thus, future phases of this work will focus on external validation using diverse datasets like TCGA-BRCA to evaluate the framework’s generalizability across different clinical populations and imaging protocols.

To enhance the clinical readiness of our interoperability module, future work will focus on formally mapping our extended schema to the official mCODE 3.0 specifications. We also plan to conduct rigorous external validation by deploying the generated FHIR bundles on a public HAPI FHIR server to test for full schema compliance and seamless data ingestion within a live Electronic Health Record (EHR) environment. Finally, prospective studies involving pathologists are critical to validate the "Human-in-the-Loop" paradigm and quantify the real-world impact of mCODE-integrated reports on diagnostic efficiency.

Ethical Statement
This study does not contain any studies with human or animal subjects performed by any of the authors. Beyond ethical compliance, the authors acknowledge that clinical deployment of this framework would require formal validation under Software as a Medical Device (SaMD) regulatory standards to ensure the safety and reliability of machine-derived evidence in oncology workflows.

Conflicts of Interest
The authors declare that they have no conflicts of interest to this work.

Data Availability Statement
The data that support the findings of this study are openly available in Breast Cancer Histopathological Database (BreakHis) (Spanhol et al., 2016) at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/.

REFERENCES
Alom, Md. R., Farid, F. A., Rahaman, M. A., Rahman, A., Debnath, T., Miah, A. S. M., & Mansor, S. (2025). An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. *Scientific Reports*, *15*(1), 17531. https://doi.org/10.1038/s41598-025-97718-5

Botsis, T., Murray, J. C., Ghanem, P., Balan, A., Kernagis, A., Hardart, K., He, T., Spiker, J., Kreimeyer, K., Tao, J., Baras, A. S., Yegnasubramanian, S., Canzoniero, J., Anagnostou, V., The Johns Hopkins Molecular Tumor Board Investigators, Pratilas, C., Xian, R. R., Gocke, C. D., Lin, M.-T., … Lehman, J. (2023). Precision Oncology Core Data Model to Support Clinical Genomics Decision Making. *JCO Clinical Cancer Informatics*, (7), e2200108. https://doi.org/10.1200/CCI.22.00108

Boumaraf, S., Liu, X., Zheng, Z., Ma, X., & Ferkous, C. (2021). A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images. *Biomedical Signal Processing and Control*, *63*, 102192. https://doi.org/10.1016/j.bspc.2020.102192

Carbonneau, M.-A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, *77*, 329–353. https://doi.org/10.1016/j.patcog.2017.10.009

Choi, J.-W., Park, S., Sohn, J., Lew, D. H., Sung, S., Choi, Y.-L., Lee, S., & Yang, E.-J. (2024). Automated digital oncologic data review for breast cancer research: AI-enabled mCODE and field-of-interest extraction framework. *Journal of Clinical Oncology*, *42*(16_suppl), e12647–e12647. https://doi.org/10.1200/JCO.2024.42.16_suppl.e12647

D'Amato, M., van der Laak, J., & Ciompi, F. (2025). *"No negatives needed": Weakly-supervised regression for interpretable tumor detection in whole-slide histopathology images* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2502.21109

Desai, A., & Mahto, R. (2025). Multi-Class Classification of Breast Cancer Subtypes Using ResNet Architectures on Histopathological Images. *Journal of Imaging*, *11*(8), 284. https://doi.org/10.3390/jimaging11080284

George, S., Campbell, N., Hillman, S. L., Harlos, E. S., Stein, D. W. J., Chan, M. Y., Chow, S. L., Elrahi, C. L., Quina, A. C., Kokolus, M. C., Casagni, M. D., Weiss, M., Anderson, D. M., Stadler, W. M., Hoff, O. C., Rivera, D. R., Kluetz, P. G., Mandrekar, S. J., & Piantadosi, S. (2025). Feasibility of structuring electronic health record data to facilitate real-world data research: ICAREdata methods applied to multicenter cancer clinical trials. *Cancer*, *131*(1), e35528. https://doi.org/10.1002/cncr.35528

Ghasemi, A., Hashtarkhani, S., Schwartz, D. L., & Shaban-Nejad, A. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. *Cancer Innovation*, *3*(5), e136. https://doi.org/10.1002/cai2.136

Leyfman, Y., Loaiza-Bonilla, A., Cortiana, V., Tuysuz, E., Kurnaz, S., Huner, O., Giritlioglu, D., Noel Meza, J. P., & Culcuoglu, C. (2025). Performance evaluation of an AI-powered system for clinical trial eligibility using mCODE data standards. *Journal of Clinical Oncology*, *43*(16_suppl). https://doi.org/10.1200/JCO.2025.43.16_suppl.e13621

Mammadov, A., Folgoc, L. L., Adam, J., Buronfosse, A., Hayem, G., Hocquet, G., & Gori, P. (2025). *Self-Supervision Enhances Instance-based Multiple Instance Learning Methods in Digital Pathology: A Benchmark Study* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2505.01109

Molefi, T., Marima, R., Demetriou, D., Basera, A., & Dlamini, Z. (2023). Employing AI-Powered Decision Support Systems in Recommending the Most Effective Therapeutic Approaches for Individual Cancer Patients: Maximising Therapeutic Efficacy. In Z. Dlamini (Ed.), *Artificial Intelligence and Precision Oncology* (pp. 259–275). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21506-3_13

Patel, V. (2024). Exploring the Potential of ResNet50 and YOLOv8 in Improving Breast Cancer Diagnosis: A Deep Learning Perspective. *International Journal of Computer Information Systems and Industrial Management Applications*, *16*, 416–431.

Sandbank, J., Bataillon, G., Nudelman, A., Krasnitsky, I., Mikulinsky, R., Bien, L., Thibault, L., Albrecht Shach, A., Sebag, G., Clark, D. P., Laifenfeld, D., Schnitt, S. J., Linhart, C., Vecsler, M., & Vincent-Salomon, A. (2022). Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. *Npj Breast Cancer*, *8*(1), 129. https://doi.org/10.1038/s41523-022-00496-w

Sandhu, A., Kim, E. J., Urueta Portillo, D., Powers, B., & Rodriguez, R. (2025). Open-source modular AI coupled with agentic AI for comprehensive breast cancer note generation and guideline-directed treatment comparison. *Journal of Clinical Oncology*, *43*(16_suppl). https://doi.org/10.1200/JCO.2025.43.16_suppl.e13685

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. https://doi.org/10.48550/ARXIV.1610.02391

Shekhar, A., & Kim, M. (2024). *Novel Development of LLM Driven mCODE Data Model for Improved Clinical Trial Matching to Enable Standardization and Interoperability in Oncology Research*. https://doi.org/10.48550/ARXIV.2410.19826

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, *63*(7), 1455–1462. https://doi.org/10.1109/TBME.2015.2496264

Terry, M., Anton, H., Casagni, M., Chamala, S., & Chen, J. L. (2023). Exchanging genomics reports between pathology labs and medical centers using the Minimal Clinical Oncology Data Element (mCODE) FHIR implementation guide. *Journal of Clinical Oncology*, *41*(16_suppl), e13575–e13575. https://doi.org/10.1200/JCO.2023.41.16_suppl.e13575

Urueta Portillo, D., Sandhu, A., Kim, E. J., Powers, B., & Rodriguez, R. (2025). Challenges in knowledge graph generation for breast cancer using open-source LLMs and the role of mCODE. *Journal of Clinical Oncology*, *43*(16_suppl). https://doi.org/10.1200/JCO.2025.43.16_suppl.e13704

Wang, L., Fu, S., Wen, A., Ruan, X., He, H., Liu, S., Moon, S., Mai, M., Riaz, I. B., Wang, N., Yang, P., Xu, H., Warner, J. L., & Liu, H. (2022). Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing. *JCO Clinical Cancer Informatics*, (6), e2200006. https://doi.org/10.1200/CCI.22.00006

Yang, E.-J., Sung, S., & Choi, Y. (2025). 481P pCR prediction in breast cancer patients using structured information extraction using mCODE KG-enhanced large language model. *Annals of Oncology*, *36*, S376–S377. https://doi.org/10.1016/j.annonc.2025.08.905

Zhang, K., Huang, T., Malin, B. A., Osterman, T., Long, Q., & Jiang, X. (2025). Introducing mCODEGPT as a zero-shot information extraction from clinical free text data tool for cancer research. *Communications Medicine*, *5*(1), 422. https://doi.org/10.1038/s43856-025-01116-x