MedGemma Chest X-Ray Diagnostic System

Enso Atlas: On-Premise Pathology Evidence Engine for Ovarian Cancer Treatment Response Prediction

MedGemma Impact Challenge Submission

Team: Enso AI
Category: Edge AI / On-Premise Deployment
Target Use Case: Treatment Response Prediction for Ovarian Cancer

---

1. Problem Statement

The Clinical Challenge

Ovarian cancer remains one of the deadliest gynecologic malignancies, with treatment decisions often relying on incomplete information. Bevacizumab, an anti-VEGF therapy, is used as a first-line treatment option for advanced ovarian cancer, but patient response varies significantly. Currently, there is no reliable histopathology-based biomarker to predict which patients will benefit from bevacizumab therapy.

Clinicians face several challenges:

- **Limited predictive markers**: Existing clinical markers inadequately stratify patients for anti-angiogenic therapy response
- **Time-intensive slide review**: Pathologists manually search for morphological features across gigapixel whole-slide images (WSIs)
- **Black-box AI concerns**: Many AI tools provide predictions without interpretable evidence, limiting clinical adoption
- **Data privacy constraints**: Healthcare institutions require on-premise solutions that keep Protected Health Information (PHI) within hospital networks

Why AI Can Help

Histopathology slides contain rich morphological information that encodes tumor microenvironment features, stromal patterns, immune infiltration, and vascular characteristics – all potentially predictive of anti-angiogenic therapy response. Recent research has demonstrated that foundation models trained on large histopathology datasets can extract meaningful representations from H&E-stained tissue, and these representations correlate with treatment outcomes.

A key study benchmarking histopathology foundation models specifically for bevacizumab response prediction from WSIs demonstrated that high-attention regions identified by AI models can serve as potential imaging biomarkers, providing a foundation for clinically deployable decision support tools.

---

2. Solution Overview

Enso Atlas Architecture

Enso Atlas is an **on-premise pathology evidence engine** designed to predict treatment response from whole-slide images while providing clinicians with interpretable, auditable evidence. The system follows three core design principles:

1. **Evidence-First**: Every prediction is accompanied by visual evidence (heatmaps, attention patches, similar cases)
2. **Local-First**: All processing occurs on-premise with no PHI transmission required
3. **Foundation-Model-Agnostic**: Modular architecture allows swapping embedding models without pipeline changes

Core Components

Path Foundation Embeddings: We utilize Google's Path Foundation model to extract 384-dimensional embeddings from 224x224 pixel H&E patches. Path Foundation is a Vision Transformer (ViT-S) specifically trained on histopathology images, providing domain-optimized representations that reduce computational requirements for downstream classifiers.

CLAM Multiple Instance Learning: Clustering-constrained Attention Multiple Instance Learning (CLAM) aggregates patch-level embeddings into slide-level predictions. The attention mechanism assigns weights to each patch, directly translating model confidence into visual evidence regions.

FAISS Similarity Search: Facebook AI Similarity Search enables rapid retrieval of morphologically similar patches from a reference cohort. This precedent-based evidence allows clinicians to compare current cases with historical outcomes.

MedGemma Report Generation: MedGemma 4B generates structured, cautious tumor board summaries grounded exclusively in the visual evidence. The model is constrained to describe morphological observations and limitations rather than prescribing treatment.

---

3. Technical Implementation

Data Pipeline: WSI to Prediction

The processing pipeline transforms gigapixel whole-slide images into interpretable predictions through four stages:

Stage 1: WSI Ingestion and Tissue Detection

Whole-slide images (SVS, NDPI, MRXS, TIFF formats) are read using OpenSlide with cuCIM fallback for GPU-accelerated I/O. Tissue regions are identified using Otsu thresholding on the grayscale thumbnail, followed by morphological operations (closing, opening) to refine the tissue mask. This approach is computationally efficient and avoids the need for a separate segmentation model.

```
WSI (20–40 GB) --> Thumbnail Extraction --> Otsu Threshold -->
Morphological Cleanup --> Tissue Mask
```

Stage 2: Patch Extraction and Embedding

A two-phase sampling strategy balances computational efficiency with diagnostic coverage:

- **Phase 1 (Coarse)**: Grid-based sampling extracts 1,000-2,000 patches at 20x magnification (224x224 pixels)
- **Phase 2 (Adaptive)**: After initial attention map generation, additional patches are sampled from high-attention regions

Each patch is embedded using Path Foundation, producing a 384-dimensional vector. Embeddings are cached as FP16 arrays (approximately 15 MB per 20,000 patches), enabling rapid reprocessing for different downstream tasks.

```
# Embedding cache storage (per slide)
embeddings: np.ndarray  # Shape: (N_patches, 384), dtype=float16
coordinates: np.ndarray  # Shape: (N_patches, 2), level-0 pixel
        coordinates
```

Stage 3: Multiple Instance Learning with CLAM

The CLAM architecture consists of:

1. **Attention Network**: Two-layer MLP with gated attention producing attention scores for each patch
2. **Instance Clustering**: Positive and negative instance classifiers for patch-level supervision
3. **Bag Classifier**: Aggregated representation classified into response categories

```
Input: {(e_i, coord_i)} for N patches
       |
       v
[Gated Attention: tanh(W_a * e) * sigmoid(W_b * e)]
       |
```

```
       v
[Attention Weights: a_i = softmax(attention_scores)]
       |
       v
[Aggregated Representation: h = sum(a_i * e_i)]
       |
       v
[Bag Classifier: p = sigmoid(W_c * h)]

Outputs:
  - p: Probability of treatment response
  - a_i: Attention weight per patch (evidence)
```

The attention weights directly correspond to the model's confidence in each region, providing interpretable evidence without post-hoc explanation methods.

Stage 4: Evidence Generation

Three evidence modalities are generated:

Heatmap Overlay: Attention weights are mapped to patch coordinates and interpolated to create a smooth overlay on the WSI thumbnail. The visualization uses a diverging colormap (blue=low attention, red=high attention) with adjustable opacity.

Top-K Evidence Patches: The 12 highest-attention patches are extracted as a clickable gallery, allowing pathologists to inspect the regions driving the prediction.

Similar Case Retrieval: FAISS performs approximate nearest-neighbor search on the mean embedding of evidence patches against a reference cohort index. Retrieved cases include their known outcomes, enabling precedent-based reasoning.

Attention Mechanism for Interpretability

The gated attention mechanism in CLAM provides mathematically grounded importance scores:

```
attention_i = softmax(W^T * tanh(V * e_i) * sigmoid(U * e_i))
```

Where: - V, U, W are learned projection matrices - The gating (sigmoid) term suppresses uninformative patches - Softmax normalization ensures attention weights sum to 1

This produces patch-level importance scores that are: - **Non-negative**: All weights >= 0 - **Normalized**: Sum to 1 across all patches - **Directly interpretable**: Higher weight = greater contribution to prediction

Report Generation Architecture

MedGemma receives a structured input bundle:

```
{
  "evidence_patches": ["<image_1>", ..., "<image_12>"],
```

```
  "prediction_score": 0.73,
  "confidence_level": "moderate",
  "task_description": "Treatment response prediction for ovarian
        cancer",
  "output_schema": { ... },
  "constraints": [
    "Describe only visible morphological features",
    "Include limitations section",
    "Do not recommend specific treatments",
    "Cite evidence patch IDs for observations"
  ]
}
```

Output is validated against a strict JSON schema requiring: - Morphology descriptions for each cited evidence patch - Explicit statement of model limitations - Suggested confirmatory tests (IHC, molecular profiling) - Safety disclaimer regarding research-use-only status

---

4. Use of HAI-DEF Models

MedGemma 4B Integration

MedGemma 4B serves as the clinical communication layer, transforming quantitative model outputs into structured, clinician-readable reports. Unlike general-purpose LLMs, MedGemma is specifically designed for medical applications, with training that emphasizes:

- Multi-patch histopathology interpretation
- Medical document understanding
- Cautious, evidence-grounded generation

Integration Architecture: MedGemma runs locally via Hugging Face Transformers with optional quantization (INT8/INT4) for reduced memory footprint. The model receives only the evidence patches and structured metadata – no patient identifiers or external context.

Grounding Strategy: To minimize hallucination risk, we implement:

1. **Input Constraints**: Only evidence patches and model outputs are provided; no access to external knowledge
2. **Schema Enforcement**: JSON output validated against predefined schema; invalid outputs trigger re-generation
3. **Prohibition List**: Explicit blocking of treatment recommendations, dosing suggestions, and definitive diagnostic statements
4. **Citation Requirements**: All morphological observations must reference specific evidence patch IDs

Local Inference: All MedGemma inference occurs on-premise. A single NVIDIA GPU with 16GB+ VRAM can run the 4B model at interactive speeds (2-20 seconds for report generation). No PHI leaves the hospital network.

Path Foundation for Histopathology Embeddings

Path Foundation provides the feature backbone for the entire pipeline:

- **Domain Specificity**: Trained on histopathology images, producing representations tuned for tissue morphology
- **Embedding Efficiency**: 384-dimensional vectors enable fast similarity search and compact storage
- **Downstream Flexibility**: Pre-computed embeddings support rapid experimentation with different classification heads

The embedding-first architecture means Path Foundation is computed once per slide, with downstream tasks (response prediction, biomarker classification, quality control) operating on cached representations.

---

5. Results and Impact

Demonstration with TCGA Ovarian Cancer Slides

Enso Atlas was validated using the publicly available ovarian bevacizumab response dataset, comprising 288 de-identified H&E whole-slide images from 78 patients. The dataset includes binary response labels (effective vs. invalid) based on RECIST criteria and clinical follow-up.

Evaluation Protocol: - 5-fold cross-validation with patient-level splits (no data leakage) - Held-out test set of 15 patients for final evaluation - Pathologist review of attention regions for biological plausibility

Processing Performance

Metric	Value
Tissue detection	2-5 seconds per slide
Patch embedding (8,000 patches)	10-30 seconds
CLAM inference	<1 second
FAISS similarity search (10 queries)	<100 milliseconds
MedGemma report generation	5-15 seconds
Total end-to-end	**30-60 seconds per slide**

These processing times are compatible with tumor board preparation workflows, where cases are typically prepared hours to days in advance.

Clinical Usability

Informal evaluation with pathologists and oncologists highlighted:

1. **Attention maps correlate with expert intuition**: High-attention regions frequently corresponded to stromal patterns, immune infiltrates, and vascular structures – features with known relevance to anti-angiogenic therapy
2. **Similar case retrieval enables precedent-based reasoning**: Clinicians valued the ability to see historical cases with known outcomes
3. **Structured reports reduce documentation burden**: MedGemma-generated summaries provided consistent formatting for tumor board packets

Potential Real-World Deployment

Enso Atlas is designed for deployment in academic medical centers and community hospitals with digital pathology infrastructure:

Hardware Requirements: - Single workstation with NVIDIA GPU (RTX 3090/4090 or A6000) - 64+ GB system RAM - 500+ GB SSD for embedding cache - Optimal: NVIDIA DGX Spark for unified memory architecture

Integration Points: - Folder-watcher mode: Drop WSI files for automated processing - REST API: Integration with existing PACS/LIS systems - Export formats: PDF reports, JSON evidence bundles, CSV for research

Privacy-Preserving Design: - Offline operation after initial installation - No outbound network connections required - All processing and storage on hospital-controlled infrastructure - Audit logging without PHI

Limitations and Future Work

Current Limitations: - Validated on single-center cohort; multi-site validation needed - Domain shift from scanner/staining variation requires calibration - Response prediction labels inherently noisy in observational data - Not validated as a medical device; research use only

Future Directions: - Stain normalization (Macenko) for cross-site robustness - Multi-task heads for additional biomarkers (HRD, TP53 status) - Integration with Enso's proprietary foundation model - Prospective clinical validation study

---

References

1. Google Health AI. "Path Foundation Model." https://developers.google.com/health-ai-developer-foundations/path-foundation
2. Google Health AI. "MedGemma 1.5 Model Card." https://developers.google.com/health-ai-developer-foundations/medgemma
3. Lu MY, et al. "Data-efficient and weakly supervised computational pathology on whole-slide images." Nature Biomedical Engineering, 2021.
4. Wang J, et al. "Histopathological whole slide image dataset for classification of treatment effectiveness to ovarian cancer." Scientific Data, 2022.
5. NVIDIA. "DGX Spark User Guide." https://docs.nvidia.com/dgx/dgx-spark/

6. Johnson J, et al. “Billion-scale similarity search with GPUs.” IEEE Transactions on Big Data, 2019.
7. Dolezal JM, et al. “Slideflow: deep learning for digital histopathology with real-time whole-slide visualization.” BMC Bioinformatics, 2024.

This work represents a research prototype for decision support. It is not intended for autonomous clinical decision-making and has not been validated as a medical device. All predictions should be interpreted by qualified healthcare professionals in the context of complete clinical information.