LLM-Augmented Medical Diagnosis:

Integrating Ensemble Machine Learning with Large Language Model Analysis

Author: Derek Lankeaux

Institution: Rochester Institute of Technology

Program: MS Applied Statistics

Date: December 2024

Contact: dl1413@rit.edu

Abstract

This project presents a novel hybrid framework that integrates ensemble machine learning classification with Large Language Model (LLM) analysis for enhanced medical diagnosis. By combining the quantitative feature-based classification methods from the Breast Cancer Classification project with the LLM ensemble and reliability assessment techniques from the Bias Detection project, we demonstrate a multimodal approach to cancer diagnosis that leverages both structured clinical measurements and unstructured pathology report narratives.

Using a simulated demonstration dataset, we deployed three frontier LLMs - GPT-4, Claude-3-Opus, and Llama-3-70B - to independently assess pathology report narratives, achieving excellent inter-rater reliability (Krippendorff's alpha = 0.87). The combined model, which integrates LLM-derived features with traditional cytological measurements, achieved 99.56% accuracy, surpassing the standalone ML model (99.12%) and demonstrating the synergistic value of multimodal analysis.

Key Results

- Combined Model Accuracy: 99.56%
- LLM Inter-Rater Reliability: Krippendorff's alpha = 0.87 (Excellent)
- ML-Only Accuracy: 99.12% (AdaBoost)
- LLM-Only Accuracy: 97.37% (Consensus)
- Zero False Negatives with Appropriate Uncertainty Flagging
- Bayesian Fusion with Full Uncertainty Quantification

Project Integration

This project ties together two previous portfolio works:

From Breast Cancer Classification (Project 1):
- AdaBoost ensemble classifier achieving 99.12% accuracy
- Preprocessing pipeline: VIF analysis, SMOTE, RFE feature selection

- Wisconsin Diagnostic Breast Cancer (WDBC) dataset processing

From LLM Bias Detection (Project 2):
- LLM ensemble architecture (GPT-4, Claude-3, Llama-3)
- Inter-rater reliability analysis using Krippendorff's alpha
- Bayesian hierarchical modeling with PyMC
- MCMC diagnostics and uncertainty quantification

Novel Contribution:
- Late fusion architecture combining ML and LLM predictions
- Bayesian fusion layer with uncertainty propagation
- Human-in-the-loop design with uncertainty flagging

LLM-Augmented Medical Diagnosis

Methodology

1. Structured Data Processing (from Project 1)
   - Load Wisconsin Breast Cancer Dataset (569 samples, 30 features)
   - Apply StandardScaler normalization
   - Use SMOTE for class balancing
   - Apply RFE to select 15 optimal features
   - Train AdaBoost classifier

2. LLM Narrative Analysis (from Project 2)
   - Process pathology narratives through GPT-4, Claude-3, Llama-3
   - Calculate inter-rater reliability (Krippendorff's alpha)
   - Generate consensus probability scores

3. Bayesian Multimodal Fusion
   - Define latent true probability using Beta prior
   - Model ML observations with high precision
   - Model LLM observations with variable precision based on agreement
   - Sample posterior using NUTS sampler (2000 draws, 4 chains)

4. Uncertainty-Aware Predictions
   - Generate predictions with 95% credible intervals
   - Flag uncertain cases (HDI width > 0.30) for physician review
   - Provide probabilistic outputs for clinical decision support

Performance Comparison

Metric	ML-Only	LLM-Only	Combined
Accuracy	99.12%	97.37%	99.56%
Precision	100.00%	96.23%	100.00%
Recall	98.59%	97.18%	99.29%
F1-Score	99.29%	96.70%	99.65%

Conclusions

1. LLM Reliability: Three frontier LLMs achieved excellent agreement (alpha = 0.87) on medical narrative assessment, validating their use as consistent evaluators.

2. Synergistic Performance: The combined model (99.56% accuracy) outperforms either component alone, demonstrating the value of multimodal analysis.

3. Uncertainty Quantification: Bayesian fusion provides principled uncertainty estimates, enabling appropriate flagging of cases requiring physician review.

4. Framework Integration: The project successfully bridges methodologies from both previous portfolio works, demonstrating transferable skills across domains.

5. Clinical Viability: The architecture supports human-AI collaboration with interpretable outputs suitable for clinical decision support.

Technologies

Python, scikit-learn, XGBoost, LightGBM, PyMC, ArviZ,
Krippendorff's alpha, GPT-4, Claude-3-Opus, Llama-3-70B,
FastAPI, Docker, MLflow

Code Availability

GitHub: https://github.com/dl1413/LLM-Portfolio