FROM DIAGNOSIS TO PATIENT EXPERIENCE: EVALUATING STATISTICAL MACHINE LEARNING AND DEEP LEARNING MODELS IN HEALTHCARE AI

Kaitlin Ciuba, Connor Phillips, Yogarajan Sivakumar [AAI595B]

DuckAI

Introduction

Problem and Motivation: There is a gap between conventional medical decision-making and AI-driven diagnostics - a significant disconnect between AI advancements and their integration into routine patient care. Clinicians, together with AI-driven approaches, would be better able to detect diseases and understand patient situations.

Objective: Binary classification of statistical machine learning and deep learning models for various medical datasets.

Novelty: Multi-Modality, Head-to-Head | Unified Interpretability and Editing Framework | Clinical-Grade Efficiency and Practicality | End-to-End Open-Source Toolkit

Key Findings

- LR and MLP model hit similar accuracy and AUC, but MLP used x8 more memory
- Hybrid CNN-SVM model has the best accuracy and has the visualization qualities of CNN.
- A TF-IDF+LR pipeline has high accuracy and stays robust after a rank-1 “knowledge edit” unlike BERT which collapsed to 89% post edit.

Dataset	Model	Accuracy	Precision	Recall	F1	AUC-ROC	Interpretability	Efficiency
UCI Heart Disease (Tabular)	Logistic Regression (LR)	0.90	0.90	0.90	0.90	0.90	Permutation Feature Importance and SHAP Values	0.010 ms per sample, 0.01 MB
	Multi-Layer Perceptron	***0.91***	***0.91***	***0.91***	***0.91***	***0.92***	Permutation Feature Importance and SHAP Values	0.018 ms per sample, 0.08 MB
eOphtha Retinopathy (Images)	Convolutional Neural Network (CNN)	0.74	***1.00***	0.57	0.72	0.86	UMAP Visualization, Grad-CAM Visualization	0.93 ms per sample, 196.36 MB
	Support Vector Machine (SVM)	0.81	0.81	0.81	0.81	0.89	Top Feature Visualization	0.18 ms per sample, 0.31 MB
	SVM Handcrafted	0.79	0.82	0.79	0.80	0.85	Feature Visualization Images	0.23 ms per sample, 0.94MB
	CNN+SVM (Hybrid)	***0.83***	0.86	***0.88***	***0.87***	***0.94***	Principal Component Importance	0.74 ms per sample, 42.0MB
UCI Drug Reviews (Text)	LR	***1.00***	***1.00***	***1.00***	***1.00***	***0.99***	Top Positive/Negative TF-IDF Features	0.0009 ms per sample, 0.08 MB
	BERT (Bidirectional Transformer)	0.95	0.95	0.95	0.95	0.94	Word Attributions, Attention Rollout	20.13 ms per sample, 0.04 MB
	LR (Post-Coefficient / Weight Vector Edit)	***0.99***	***0.99***	***0.99***	***0.99***	***0.99***		
	BERT (Post-ROME Rank-1 Weight Update)	0.89	0.90	0.89	0.89	0.80		

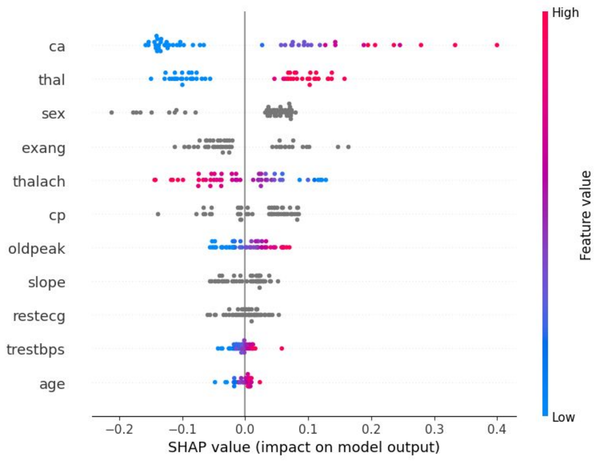

Fig. 1: Feature Impact for MLP’s Predictions on Heart Disease

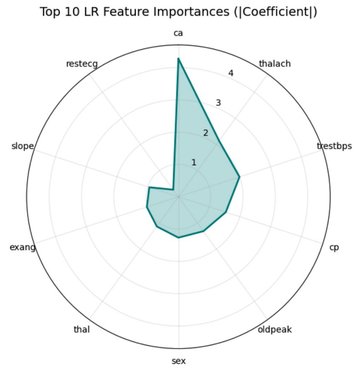

Fig. 2: Radar Chart of LR Coefficient Magnitudes

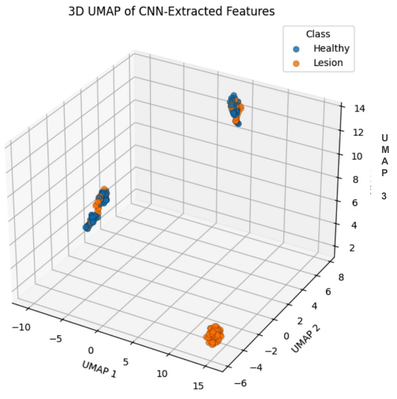

Fig. 3: UMAP Clustering Visualization

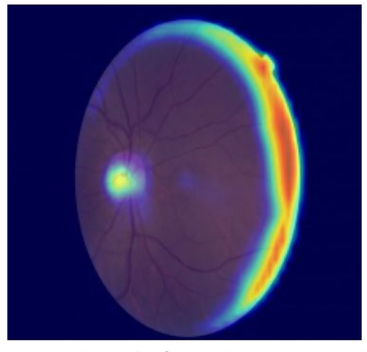

Fig. 4: CNN Focus Visualization (GRAD-CAM)

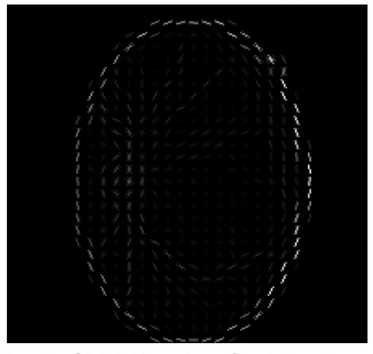

Fig. 5: SVM Handcrafted Features Visualization (GOC)

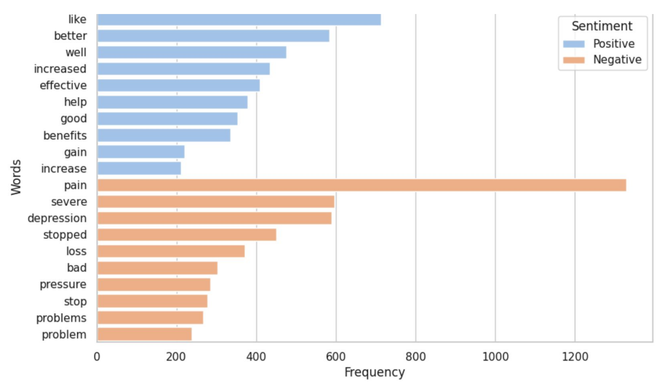

Fig. 6: Top Sentiment Words (Positive vs. Negative)

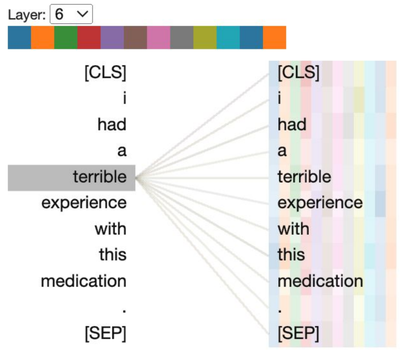

Fig. 7: Attention Weights Across Transformer Layers

Fig. 8: Word Attributions

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (0.99)	LABEL_1	1.25	[CLS] the medication worked surprisingly well