Enhancing Clinical Decision Making with Interpretable AI

Author: **Anish Narain** | Supervisor: **Dr Sonali Parbhoo** | **AI for Actional Impact Lab**

IMPERIAL

Department of Electrical and Electronic Engineering

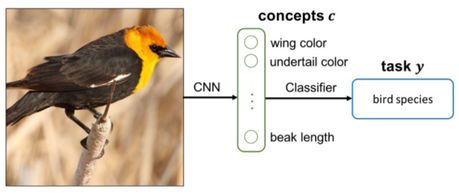
Background

Large Language Model (LLM)

+ Impressive performance on benchmark medical exams
+ Demonstrates strong reasoning on a variety of contextual data

- Black-box
- User cannot inject new knowledge
- Can misbehave and hallucinate

Concept Bottleneck Model (CBM)



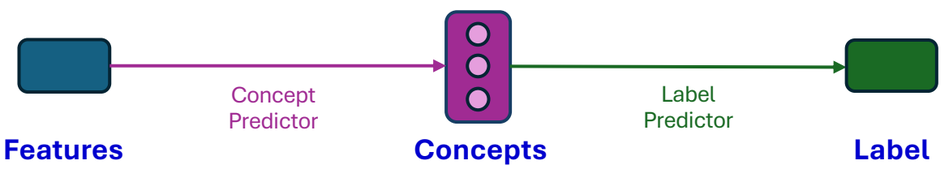
+ Interpretable & intervenable
- Limited by leakage & quality of training data

Contributions

1. Implemented a **Mortality Prediction CBM** trained on EHR data with 83% accuracy, comparable to regression baseline, with superior interpretability.
2. Implemented an **ARDS Identification CBM** trained on EHR data with 68% accuracy, outperforming regression baseline by 7%.
3. **Augmented ARDS CBM** with additional concepts from clinical notes **using LLM**. The accuracy improved by up to 12%, and the likelihood of leakage was reduced.

1. Mortality Prediction CBM

Mortality (Patient Death)



6 SOFA Organ Components
Respiratory, Liver, Cardiovascular, Coagulation, CNS, Renal

Lab Measurements
Bilirubin, Albumin, Aniongap

Moderate/Severe SOFA Organ Failure

Potential Conditions

High/Low Lab Measurements

Mortality

Septic shock, ARDS, Hepatorenal Syndrome, COD, MODS, CNS and Renal Failure, Liver and Coagulation Failure

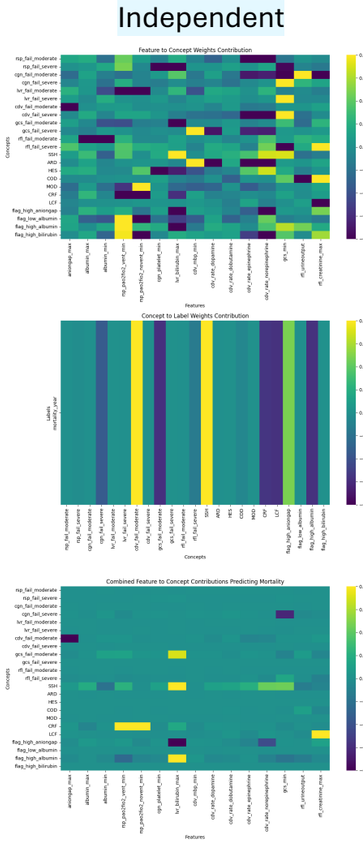
Evaluation Metrics

Concept Predictor's Model	Accuracy	AUC	Precision	Recall	F1 Score
Independent	0.99	0.87	0.83	0.74	0.76
Sequential	0.99	0.87	0.83	0.74	0.87
Joint	0.98	0.86	0.72	0.73	0.72
Independent w/ Early Stopping	0.98	0.81	0.78	0.63	0.68
Sequential w/ Early Stopping	0.99	0.90	0.85	0.80	0.82
Joint w/ Early Stopping	0.96	0.75	0.74	0.55	0.60
Independent w/ L2 Regularisation	0.99	0.87	0.85	0.74	0.78
Sequential w/ L2 Regularisation	0.99	0.87	0.85	0.74	0.78
Joint w/ L2 Regularisation	0.98	0.84	0.83	0.69	0.73

Label Predictor's Model	Accuracy	AUC	Precision	Recall	F1 Score
Independent	0.82	0.62	0.71	0.27	0.40
Sequential	0.82	0.63	0.75	0.28	0.40
Joint	0.82	0.64	0.70	0.33	0.45
Independent w/ Early Stopping	0.82	0.63	0.77	0.28	0.41
Sequential w/ Early Stopping	0.82	0.63	0.74	0.29	0.42
Joint w/ Early Stopping	0.83	0.65	0.76	0.32	0.45
Independent w/ L2 Regularisation	0.82	0.63	0.79	0.27	0.41
Sequential w/ L2 Regularisation	0.83	0.64	0.78	0.30	0.44
Joint w/ L2 Regularisation	0.83	0.66	0.70	0.37	0.48

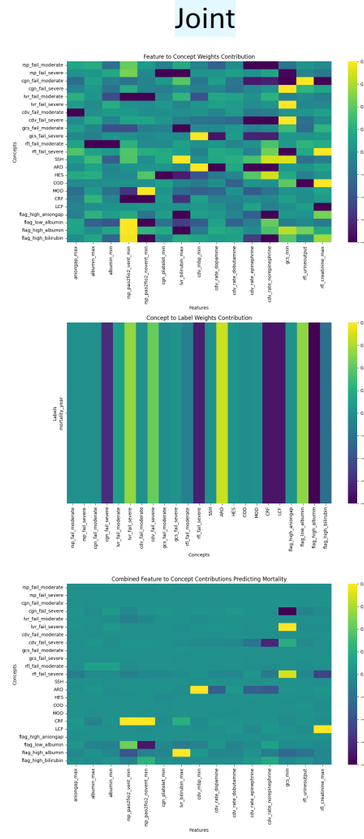
Heat Maps

Independent



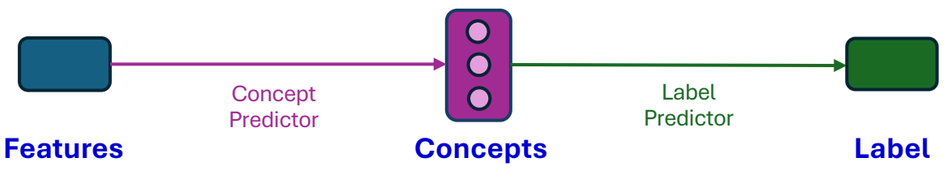
Sequential

Joint



2. ARDS Identification CBM

ARDS (Acute Respiratory Distress Syndrome)



3 SOFA Organ Components
Respiratory, CNS, Renal

Sepsis & Shock Lab Values

Ventilation Status

Pre-existing Respiratory Conditions

Moderate/Severe SOFA Organ Failure

Sepsis or Shock Detected

Severity of Ventilation

Severity of Pre-existing Conditions

ARDS Diagnosed by Dr Marshall

- A cohort of patients potentially having ARDS was screened.
- After manual review of EHR and notes, a final diagnosis was made.

Evaluation Metrics

Concept Predictor's Model	Accuracy	AUC	Precision	Recall	F1 Score
Independent	0.96	0.82	0.87	0.80	0.83
Sequential	0.96	0.82	0.87	0.80	0.83
Joint	0.94	0.82	0.79	0.76	0.76
Independent w/ Early Stopping	0.96	0.82	0.88	0.81	0.83
Sequential w/ Early Stopping	0.96	0.82	0.87	0.81	0.83
Joint w/ Early Stopping	0.89	0.62	0.68	0.60	0.62
Independent w/ L2 Regularisation	0.96	0.83	0.88	0.81	0.83
Sequential w/ L2 Regularisation	0.96	0.83	0.88	0.81	0.83
Joint w/ L2 Regularisation	0.96	0.82	0.87	0.80	0.82

Label Predictor's Model	Accuracy	AUC	Precision	Recall	F1 Score
Independent	0.57	0.58	0.52	0.75	0.61
Sequential	0.55	0.56	0.50	0.71	0.59
Joint	0.61	0.62	0.57	0.65	0.61
Independent w/ Early Stopping	0.62	0.62	0.58	0.63	0.60
Sequential w/ Early Stopping	0.61	0.60	0.59	0.49	0.54
Joint w/ Early Stopping	0.68	0.68	0.65	0.65	0.65
Independent w/ L2 Regularisation	0.63	0.63	0.59	0.63	0.61
Sequential w/ L2 Regularisation	0.60	0.59	0.56	0.57	0.57
Joint w/ L2 Regularisation	0.67	0.67	0.63	0.69	0.66

3. Augmenting ARDS CBM using LLM

LLM Concept Generation

```
template=(
    "Context: You are a clinician receiving chunks of radiology reports for
        patients in an ICU. Please do the reviewing as quickly as possible.\n"
    "Task: Determine if the patient suffered from bilateral infiltrates.\n"
    "Instructions: Answer with 'Yes' or 'No'. If there is not enough
        information, answer 'No'.\n"
    "Discharge Text:\n{radiology_texts}\n\n"
    "Query: Does the chunk of text mention that the patient suffered from
        bilateral infiltrates? Answer strictly in 'Yes' or 'No'.  Then provide
        a reason for your response."
),
input_variables=["radiology_texts"]
```

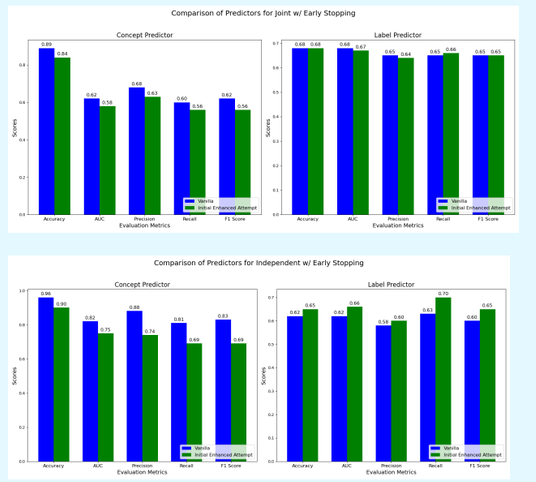
Using **Meta's Llama3 model** on discharge summaries, radiology reports, and ECG studies.

Clinical Label	Average Accuracy (%)
Mention of ARDS	93.5
Aspiration	87.4
Bilateral Infiltrates	71.7
Cardiac Arrest	80.3
Cardiac Failure	81.2
Pancreatitis	74.6
Pneumonia	92.7
TRALI	96.3

Average accuracy of clinical labels from random sampling validation.

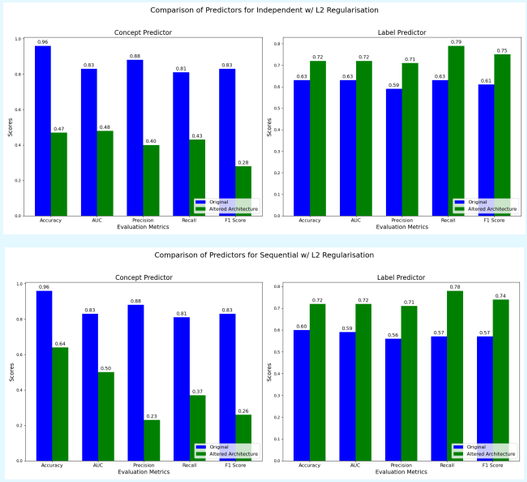
Enhanced CBM Architecture

Initial Attempt



Model Type	Configuration	Completeness	I(y; c)	I(y; c, x)
Vanilla	Joint w/ Early Stopping	0.057	0.040	0.69
	Joint w/ L2 Regularisation	0.075	0.052	0.684
Enhanced	Joint w/ Early Stopping	0.059	0.041	0.69
	Joint w/ L2 Regularisation	0.075	0.052	0.68

Altered Architecture



Model Type	Configuration	Completeness	I(y; c)	I(y; c, x)
Vanilla	Independent w/ L2 Regularisation	0.026	0.017	0.661
	Sequential w/ L2 Regularisation	0.025	0.017	0.682
Enhanced	Independent w/ Altered Architecture	0.18	0.13	0.68
	Sequential w/ Altered Architecture	0.73	0.1364	0.1864

References

[1] (Logo) "Imperial Brand Project," *Imperial College London*. https://www.imperial.ac.uk/communications/about-us/projects/imperial-brand-project/

[2] (Background) Bird Identification CBM Image from P. W. Koh, T. Nguyen, Y. S. Tang, et al., "Concept bottleneck models," in International conference on machine learning, PMLR, 2020, pp. 5338–5348.