Improving ARDS Diagnosis Through Context-Aware Concept Bottleneck Models

Authors: Anish Narain, Ritam Majumdar, Nikita Narayanan, Dominic C Marshall, Sonali Parbhoo
Contact: s.parbhoo@imperial.ac.uk

IMPERIAL

AI4AI
AI for Actional Impact Lab

Project GitHub

1. Overview

Problem

Clinical Challenges in ARDS Diagnosis

- Requires various data sources (vitals, labs, imaging, clinical notes)
- Inconsistencies in documentation
- Variability in clinical judgment
- Absence of single gold-standard test

Traditional ML Models

+ Shown potential for automating ARDS detection using **structured data**

- Limited interpretability

Concept Bottleneck Models

+ **Interpretable** and clinical **intervention** possible

- Risk of **concept leakage**

Why a Retrospective ARDS Study?

- **Acute Respiratory Distress Syndrome (ARDS)** is often **missed at the bedside** and **poorly documented** when identified, limiting accurate data collection.
- Retrospective labelling enables the **creation of large, accurate ARDS patient cohorts** from existing ICU databases.
- This method facilitates observational studies to **identify patient subtypes** and analyze **varied treatment responses** in ARDS.

Solution
Context-Aware CBMs

2. Context-Aware CBM

Problem Setup

$\left(x^{(i)}, y^{(i)}, c_j^{(i)}\right)_{i=1}^{n}$	Training set of input features x, target labels y, and vanilla concepts c
$f(g(x))$	CBM where g maps input x into concept space and f maps concept into a final prediction. Trained CBM is represented using $\hat{f}$ and $\hat{g}$
L_Y	Loss function that measures difference between predicted and true target labels y
L_{C_j}	Loss function that measures difference between predicted and true j-th concepts

Context-Aware CBM Loss Function

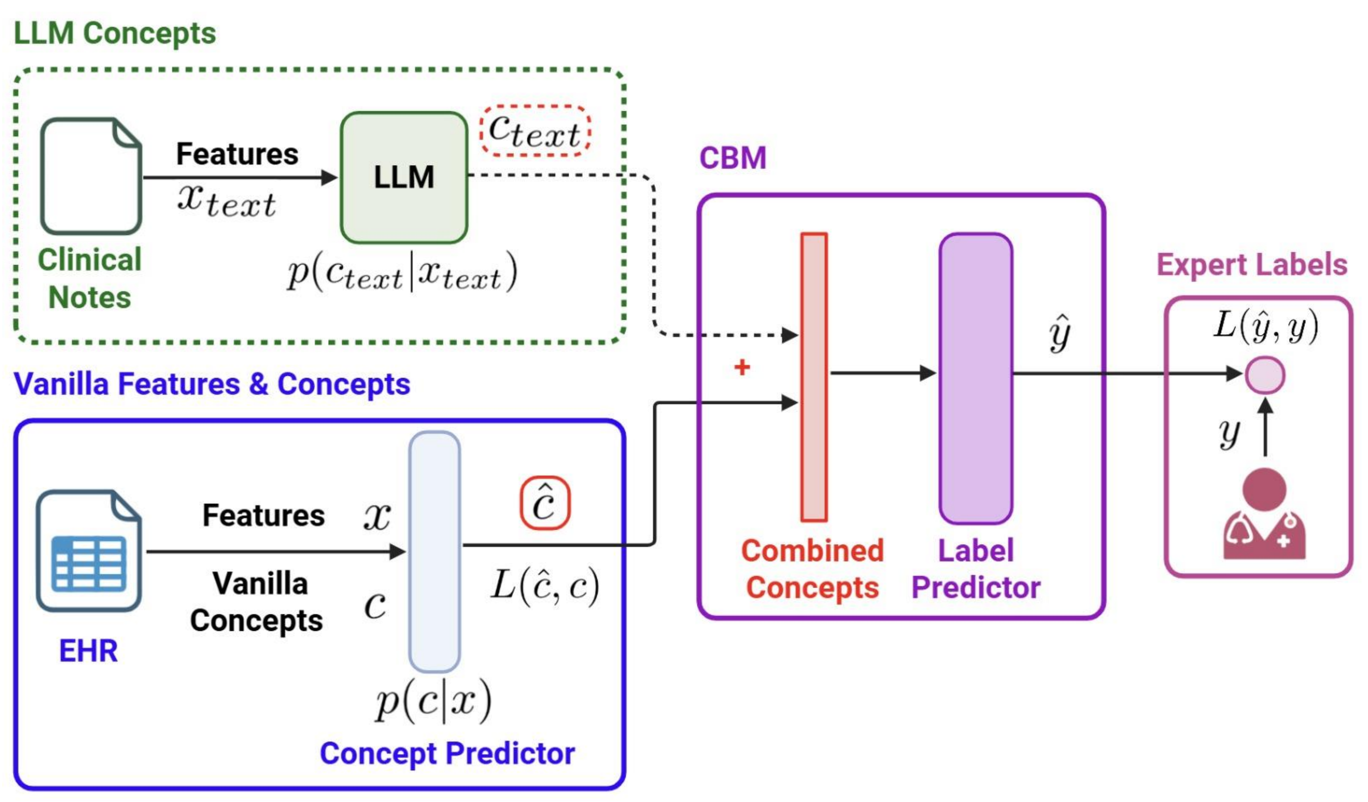
In Context-Aware CBMs, the initial predicted concepts come from distribution $p(c|x)$ and the LLM concepts come from a different distribution $p(c_{text}, x_{text})$. The loss function thus becomes:

$$\hat{g}, \hat{f} = \arg\min_{f,g} \sum_i \left[L_Y\left(f\left(g(x^{(i)}), c_{text}\right); y^{(i)}\right) + \sum_j \lambda L_{C_j}\left(g(x^{(i)}); c^{(i)}\right) \right]$$

LLM Prompt Template

```
prompt_template=(
    "Context: You are a clinician receiving chunks of clinical text for patients in an ICU. Please do the
    reviewing as quickly as possible."
    "Task: Determine if the patient suffered from pancreatitis."
    "Instructions: Answer with 'Yes' or 'No'. If there is not enough information, answer 'No'."
    "Discharge Text:{discharge_text}"
    "Query: Does the chunk of text mention that the patient suffered from pancreatitis? Answer strictly in
    'Yes' or 'No'."
),
input_variables=["discharge_text"]
```

Training Pipeline for Context-Aware CBMs



3. Cohort Selection and Dataset

Cohort Selection Criteria

- Chose adult patients from MIMIC IV dataset who were not pregnant and not on ECMO
- The patients also met respiratory criteria: PF ratio < 300 with PEEP ≥ 5 cm H2O for three consecutive days, or two days if the patient died on day 3

An expert clinician then reviewed clinical records and assigned each patient a **label**. Of the 1953 patients reviewed:

- 1030 (52.7%) labelled **positive** for ARDS
- 923 (47.3%) labelled **negative** for ARDS

Features

Vitals used to calculate SOFA score for 4 organ systems, vasopressor support, mechanical ventilation duration, and pre-existing respiratory comorbidities

Vanilla Concepts From Tabular EHR Data

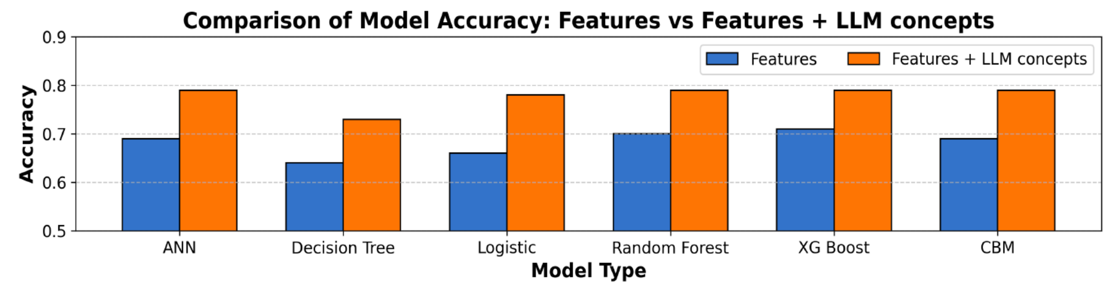
Average and max SOFA scores for respiratory, cardiovascular, renal and central nervous systems. Also, severity of pre-existing respiratory comorbidities

LLM Concepts From Clinical Text

Pneumonia, aspiration, pancreatitis, cardiac arrest, TRALI, bilateral infiltrates, cardiac failure, overall clinical impression of ARDS

4. Results

1. Context-Aware models augmented with LLM concepts show an **8-10% improvement over traditional models** across all metrics



__Result #1:__ Comparative model performance between Features and Features augmented with LLM concepts.

2. Context-Aware CBMs **outperform vanilla CBMs by 10-20%** across all metrics, indicating that LLMs extract important concepts relevant for final ARDS prediction

3. Context-Aware CBMs product concepts which have **higher mutual information**, indicating **leakage reduction**

4. **Interventions** on misclassified concepts further **boost the performance** of Context-Aware CBMs by **12-20%**

Model Type	Acc.	Prec.	Recall	F1	M1
Vanilla CBM	0.69	0.71	0.72	0.71	0.08
Context-Aware CBM	0.79	0.76	0.88	0.81	0.21
Vanilla CBM + Vanilla concept interventions	0.85	0.88	0.86	0.90	0.78
Context-Aware CBM + Vanilla concept interventions	0.87	0.88	0.91	0.92	0.84
Context-Aware CBM + LLM concept interventions	0.94	0.92	1.00	0.93	0.92
Context-Aware CBM + (LLM + Vanilla) concept int.	0.96	1.00	0.93	0.97	0.96

__Result #2,3,4:__ Performance of Vanilla and Context-Aware CBMs augmented with interventions over misclassified patients

5. Context-Aware CBMs **improve prediction** of concepts **critical to the outcome**, thereby **reducing leakage**

6. **Interventions** based on **correlated concepts** are **more effective** than interventions on individual concepts

7. Context-aware CBMs **outperform** vanilla CBMs on **out-of-distribution patients**

5. Case Studies

Context-aware CBMs are **interpretable** and **intervenable**. These properties allow us to identify key diagnostic factors in patient cases and explore how intervening on the concepts affect the diagnosis.

Study 1: Intervention to Correct False Negatives

- **Problem:** Vanilla CBM (using structured EHR data only) missed ARDS in two patients
- **Fix:** The **Context-Aware CBM**, with LLM access to clinical notes, identifies **key textual evidence** (e.g., "bilateral infiltrates," "pneumonia," absence of "cardiac failure") that points to ARDS.
- **Result:** The diagnosis was corrected to ARDS-positive.

Study 2: Intervention to Correct LLM-Induced Errors

- **Problem:** Context-Aware CBM wrongly **over-relies on LLM-inferred concepts** that are not actually present or relevant.
  - **Patient 1:** LLM erroneously picks up "pneumonia/ARDS" → false positive.
  - **Patient 2:** LLM invents a "cardiac arrest" concept → false negative.
- **Fix:** Manual **intervention at the concept level** (adjusting or removing erroneous LLM concepts).
- **Result:** Restored correct predictions made by the vanilla CBM.

Study 3: Concept-Level Debugging

- **Problem:** Both vanilla and Context-Aware CBMs misclassify two patients—one ARDS-positive and one ARDS-negative—due to **inaccurate intermediate concept predictions** (e.g., wrong morbidity scores, missing comorbidities).
- **Fix:** The **interpretability of CBMs** was used to identify which **specific intermediate concepts** (e.g., morbidity, cardiac failure) are incorrect.
- **Result:** Intervening on those concepts **corrected the predictions**.

6. Conclusion

Paper Contributions

- [x] 1. Proposed a general framework for **enhancing CBMs** using context from **unstructured data**, applicable to clinical use cases requiring multi-modal reasoning.
- [x] 2. Achieved an **8-10% improvement** in **retrospective ARDS diagnosis**.
- [x] 3. Enriched concepts with LLM-derived context. This **increased concept completeness**, thereby **mitigating concept leakage**.
- [x] 4. Enabled transparent concept-level reasoning, allowing for **interventions to correct errors**, thus improving reliability for clinical use.

Paper Limitations

- **Delayed Diagnosis**: The current study relied on retrospective access to complete patient data. In practice, ARDS diagnosis requires rapid integration of diverse data soon after onset, so future models must use data available within a specific, real-time window.
- **Noisy LLM Concepts**: Concepts generated by large language models may introduce irrelevant or misleading signals into predictions. Incorporating human oversight could help reduce errors before clinical application.