Applied Deep Learning for Early Mortality Prediction in ICU Heart Failure Patients: A Summary Report

Jacob Scott
Student, MS Data Science
Indiana University
Bloomington, Indiana
scotjaco@iu.edu

***Abstract*—Heart failure (HF) is a major contributor to in-hospital mortality among ICU patients, highlighting the potential need for effective early warning systems. This applied body of work utilizes the MIMIC-IV database to predict in-hospital mortality for ICU patients diagnosed with HF, using only the first 48 hours of data post-ICU-admission to simulate an early warning model. Three datasets were constructed: static patient information, temporal events data (including labs, weight, and vasoactive agents), and bedside monitoring vitals aggregated at 5-minute intervals. A machine learning approach (logistic regression) and various implementations of a deep learning approach, a multi-layer perceptron (MLP), were implemented and evaluated across different data integration levels, attempting to build a multi-modal model utilizing all three datasets. Model performance was assessed using metrics like ROC-AUC and F1-score. Results indicate that logistic regression performs well initially and on the isolated datasets. However, the deep learning approaches became much more relevant when integrating multiple data sources. The highest performing model was the MLP utilizing all three sources. This demonstrates the potential of deep learning models in enhancing early mortality prediction for HF patients in the ICU, particularly with large, multi-modal datasets. These preliminary findings may support the development of scalable, data-driven early warning systems aimed at improving clinical decision-making and patient outcomes in critical care settings.**

I. Important Notes

A. Limited Scope

Due to the project's scaled-back scope, this summary report does not include an extensive background literature review, though the data and its seminal paper and other works that influenced this research are cited throughout.

Rather, this paper focuses directly on the dataset creation, data processing, and modeling methodologies employed to predict in-hospital mortality for ICU patients with a heart failure (HF) diagnosis.

II. Introduction

A. Objectives

The primary objective of this work is to develop an exploratory deep learning model aimed at serving as an early-warning indicator for in-hospital mortality for ICU patients diagnosed with Heart Failure (HF). Utilizing the initial 48 hours of patient demographic and clinical data, this model will determine a positive or negative indication for mortality later in the hospital stay.

This work will also evaluate the impact of progressively integrating the three disparate data sources extracted for this model (static demographics, aggregated event data, and flattened vitals) on mortality prediction performance.

B. MIMIC-IV Database Overview

The Medical Information Mart for Intensive Care IV (MIMIC-IV) is a publicly accessible electronic health record (EHR) database that encompasses comprehensive clinical data from patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Covering admissions from 2008 to 2019, MIMIC-IV includes detailed information such as demographics, vital signs, laboratory test results, medications, and clinical notes. This dataset supports a wide range of research applications, including epidemiological studies, clinical decision support development, and machine learning model training. MIMIC-IV enhances its predecessors by adopting a more modular data structure and facilitating the integration of new and disparate data sources. [1]

The version used for this research is version 2.2, accessed via Google BigQuery. [2]

III. Data Extraction and Integration

The datasets used in this body of work were created by querying the MIMIC-IV database via Google BigQuery, containing the de-identified health records of ICU patients as described above. Data extraction was tailored to the research objectives, focusing on ICU patients diagnosed with heart failure via ICD-9 and ICD-10 codes and capturing relevant information within the first 48 hours of ICU admission to support the objective of early mortality prediction.

Three SQL queries were developed to extract data from different perspectives:

A. Queries

1) Static Patient Data

This query retrieved demographic and static clinical attributes that provide foundational context about the patient's baseline characteristics and health history, forming the static dataset.

2) Bedside Monitoring Vitals

This query focused on high-frequency time-series data, such as heart rate, blood pressure, respiratory rate, and oxygen saturation, collected through bedside monitoring equipment. The data were subsequently aggregated into 5-minute intervals to reduce granularity while retaining meaningful trends and variability. It is worth noting that HF is a slow-moving condition in terms of clinical monitoring data. This dataset provides continuous insights into a patient's current physiological state.

3) Temporal Events

The temporal events query extracted time-stamped event data, such as laboratory test results, weight measurements, and vasoactive agent administrations. These features represent key clinical interventions and measurements recorded during the ICU stay, aggregated and aligned to complement the static data. This dataset captures critical changes in a patient's condition over time.

B. Data Integration

The extracted datasets were created with the ability to be joined using `stay_id`, a unique identifier for ICU admissions, to create a methodology for analysis. Temporal datasets (events and vitals) were aligned to ensure consistency within the first 48 hours of ICU admission, and static patient data provided a stable reference across all analyses.

C. Features

The features encompassed by each dataset are listed in the following tables below:

TABLE I: Static Patient Data Features

Feature	Description
stay_id	ICU stay identifier (unique across all records)
gender	Patient gender
age	Patient age at admission
race	Patient self-reported race
marital_status	Patient self-reported marital status
admission_type	Admission class (e.g. Emergency)
admission_location	Source of admission (e.g. Referral)
hours_to_icu	Time in hours after hosp. adm. to ICU adm.
prior_hospital_admissions	Total admissions. before current
prior_icu_admissions	Total ICU admissions before current
ace_inhibitors_flag	Prescription med. indicator
arbs_flag	Prescription med. indicator
arnis_flag	Prescription med. indicator
beta_blockers_flag	Prescription med. indicator
loop_diuretics_flag	Prescription med. indicator
thiazide_diuretics_flag	Prescription med. indicator
sglt2_inhibitors_flag	Prescription med. indicator
mras_flag	Prescription med. indicator
hospital_expire_flag	Target feature - Mortality indicator

TABLE II: Bedside Monitoring Vitals Features

Feature	Description
stay_id	ICU stay identifier (unique across all records)
charttime	Timestamp of measurement
heart_rate	Heart rate of patient
systolic_bp	Systolic blood pressure of patient
diastolic_bp	Diastolic blood pressure of patient
mean_bp	Mean of blood pressure measurements
respiratory_rate	Respiratory rate of patient
spo2	Blood oxygen saturation of patient

TABLE III: Temporal Events Features

Feature	Description
stay_id	ICU stay identifier (unique across all records)
timestamp	Timestamp of event (Chart time, lab time, etc.)
event_type	One of: lab, weight, or vasoactive agent
variable_name	Specific measurement (e.g. Phenylephrine)
value	Float value of the measurement

IV. Data Processing

The data processing workflow focused on preparing the three extracted datasets—static patient data, time-series

vitals data, and temporal events data—for predictive modeling. The goal was to ensure quality, consistency, and alignment across the datasets while handling outliers, missing values, and feature engineering.

A. Static Patient Data

Static patient data included demographics, comorbidities, and admission details. Patients who spent more than 30 days in a standard ward before ICU admission were excluded as outliers. Only `stay_ids` shared across all datasets were retained. Features such as `gender` and `race` were simplified and encoded, with race categories grouped into the broader categories defined within the data, such as "White," "Black/ African," and "Hispanic/Latino." One-hot encoding was applied to categorical features, and boolean columns were converted to binary integers. The target variable, `hospital_expire_flag`, showed a 12% mortality rate, highlighting a significant class imbalance, which will be covered in greater detail further on.

B. Bedside Monitoring Vitals

Vitals were resampled to 5-minute intervals and interpolated to handle missing values. Patients with fewer than 6 hours of data (72 records) or entirely missing data for any vital sign were excluded. Temperature was dropped due to inconsistent availability. The cleaned dataset retained sufficient coverage of key vitals, such as blood pressure, respiratory rate, and oxygen saturation.

C. Temporal Events

Event-based clinical data were aggregated using statistical functions (mean, median, min, max) for each variable grouped by stay_id. Features with a missing ratio exceeding the majority class ratio were dropped, while the remaining features were imputed using medians and enhanced with binary missingness indicators, as the lack of information can be equally informative in some cases. These missingness indicators were inspired by the indicator vector for padded values described by research of deep learning applications for long-term HF patient mortality prediction for MIMIC-III. [3]

V. Modeling Methodology

A. Models Implemented

Two primary models were implemented at each level of data integration of the three datasets:

1) Logistic Regression (Baseline)
Used as a straightforward, linear baseline, including class-weight adjustments for imbalance.

2) Multilayer Perceptron (MLP) Neural Network: A feedforward neural network with several dense layers, batch normalization, dropout, and a focal loss function to address class imbalance.

B. Addressing Class Imbalance

As noted, the dataset exhibited a severe class imbalance, with the positive (mortality) class comprising only 12.8% of the population. This skewed distribution can lead predictive models to inordinately favor the majority class, resulting in low recall for the minority outcome and in this case, underestimation of the in-hospital mortality events. The imbalance influenced both modeling decisions and evaluation strategies. For instance, simple accuracy metrics became less informative, prompting the use of performance measures sensitive to minority class detection, such as minority F1-score and ROC-AUC. Additionally, specialized modeling techniques were employed to mitigate this imbalance. Logistic regression models incorporated class-weight adjustments, and the MLP neural network implemented Focal Loss [4] to emphasize harder-to-classify minority examples. These choices were made to try to ensure that improvements reflected an actual increase in sensitivity rather than a mere optimization on the overwhelmingly dominant class.

C. On Focal Loss

Focal Loss is a modified form of cross-entropy loss designed to address class imbalance by downweighting easy, correctly classified examples. It introduces a focusing parameter γ (gamma) that adjusts the rate at which easy examples are discounted, and an α (alpha) parameter to weight classes differently. This encourages the model to pay more attention to hard-to-classify minority instances. [4]

D. Data Integration Levels

Static Only: The model inputs consist solely of patient-level static features (demographics, admission details, etc.).

Static + Aggregated Events: In addition to the static features, this level incorporates aggregated event-based features. This merging includes statistical summaries (mean, median, etc.) of temporal event data to enhance static inputs.

Full Integration (Static + Aggregated Events + Flattened Vitals): The static and event-based features are further combined with flattened, time-aggregated vital sign statistics. This produces a richer feature set blending demographic, event, and physiological signals.

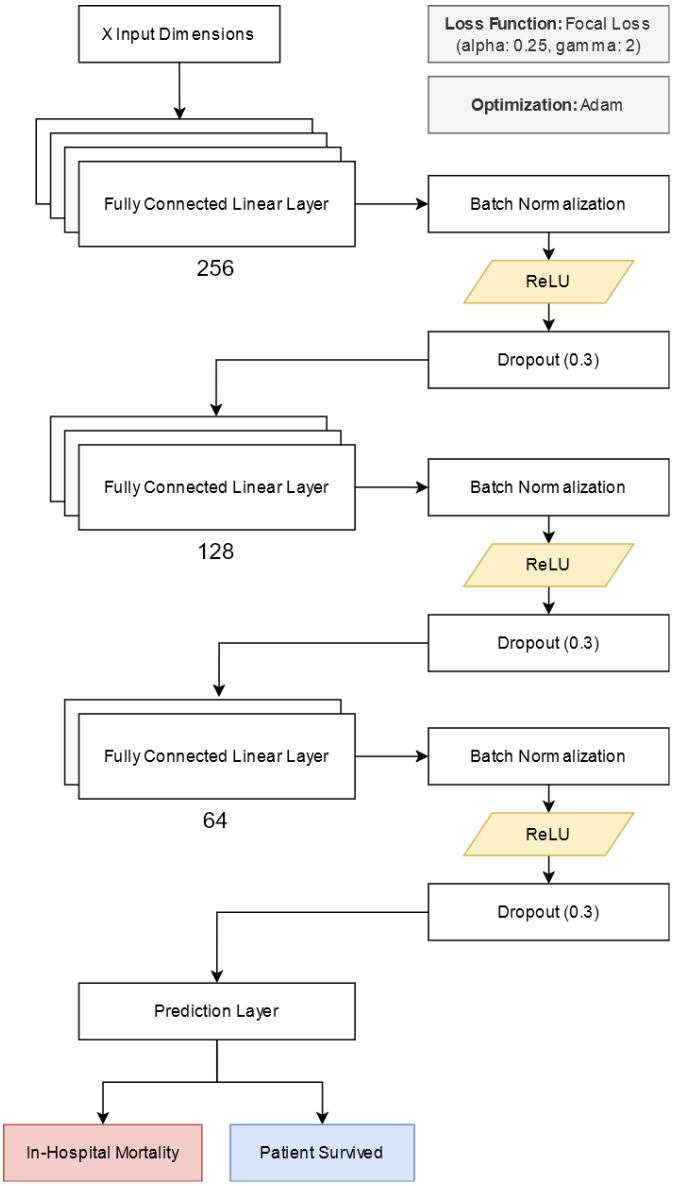

Fig. 1: Neural network architecture of the Full Integration MLP model.

E. Training and Evaluation

All models were trained using a standard train-test split. As all features were converted to numeric format, they are all scaled with `StandardScaler`, ensuring that scaling is performed after the split to mitigate potential for data leakage. The logistic regression provided a baseline for each integration level. The MLP network, trained with the Adam optimizer and a chosen learning rate (scheduled LR for the full integration level), incorporated Focal Loss to more effectively handle the severe class imbalance. Evaluation metrics included ROC-AUC, F1-score, and confusion matrices, focusing on improvements from baseline (logistic regression) to the more flexible MLP model as additional data integration strategies were iteratively introduced.

VI. Results

A. Performance Comparison

Static Only:

- Logistic Regression: Achieved a moderate balance (minority-class F1 0.34) and ROC-AUC of 0.7573.
- MLP: Displayed high specificity but poor sensitivity, resulting in a very low minority-class F1 (0.04). ROC-AUC was 0.7494.

Static + Aggregated Events:

- Logistic Regression: Substantial improvement in minority-class recognition (F1 0.47) and ROC-AUC of 0.8516.
- MLP: Again showed strong majority-class performance but limited gains in minority detection (ROC-AUC: 0.7427).

Full Integration (Static + Aggregated Events + Flattened Vitals):

- Logistic Regression: Further enhanced minority-class detection (F1 0.49) and an improved ROC-AUC of 0.8730.
- MLP (0.5 Threshold): Markedly improved performance, achieving a minority-class F1 of 0.50 and ROC-AUC of 0.8786.
- MLP (Optimized Threshold): With threshold tuning, the MLP demonstrated its best balance, improving the minority-class F1 to 0.58 while maintaining a strong ROC-AUC of 0.8786. See Fig. 2 for ROC curve and Fig. 3 for a threshold classification comparison via confusion matrices (next page).

More complete performance details can be found in Table IV.

Results continue on the next page.

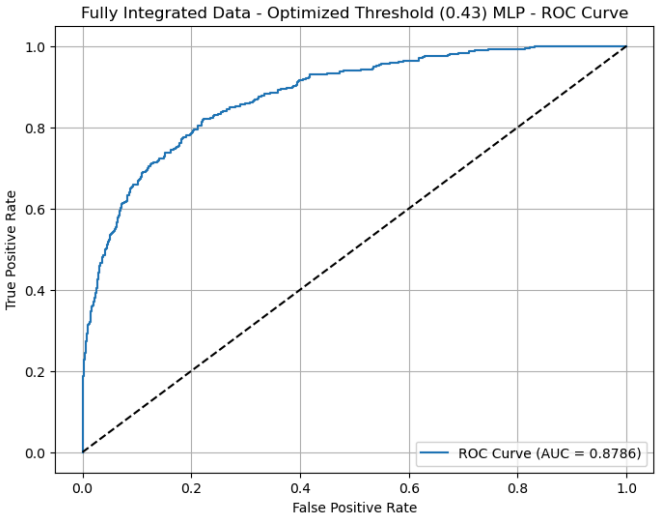

Fig. 2: The receiver operating characteristic (ROC) curve for the MLP tuned for the optimal threshold via precision-recall performance, showing a high degree of distinguishability between positive and negative classes.

B. Key Findings

- Adding event-based and physiological features (full integration) consistently improved baseline logistic regression and MLP models' ability to identify the minority class.
- While the MLP struggled with minority detection in less integrated scenarios, fully integrated data and threshold calibration yielded the strongest overall performance. It is important to note that the lower integration levels did not receive the same degree of attention to MLP minority classification performance, merely overall performance.
- The best result was achieved by the fully integrated MLP with an optimized threshold, offering a well-rounded improvement in both discriminatory power (high ROC-AUC) and minority-class F1-score.
- These findings highlight the relevance of additional feature integration and post-hoc calibration in addressing class imbalance challenges.

VII. Discussion

A. On the Final Model (Optimized Full-Integration MLP)

This model is the best of the attempts at overcoming a significant challenge (RE: class imbalance). It correctly identifies the majority of patients at risk of mortality, capturing about 6 in 10 of those who will not survive while maintaining strong accuracy overall (89%). In practice, a model such as this, developed and deployed with careful consideration, could offer clinicians a valuable early-warning tool, highlighting most high-risk patients early on with relatively few unnecessary alarms.

B. Balancing Clinical Relevance and Predictive Performance

While these results indicate that logistic regression provides a strong baseline and that deep learning models can excel when leveraging the more extensive multimodal data, it's crucial to interpret these outcomes with patient care in mind. Traditional metrics like ROC-AUC and overall accuracy confirm a model's capacity to distinguish between classes, but they do not ensure clinically meaningful performance by default. When high specificity comes at the cost of missing a large fraction of critically ill patients, the model's predictive power becomes less useful in a real-world setting. For example, the left panel in Fig. 3 shows the final, full-integration MLP model at its default classification threshold of 0.5. Of 367 true mortality cases, only 144 were detected at this threshold for a recall of 0.39. If 61% of mortality cases are not flagged by the model, then even a seemingly strong ROC-AUC has limited practical value.

C. Importance of Sensitivity and Specificity

Sensitivity and specificity are essential metrics in this clinical context. Sensitivity (true positive rate) measures how well the model identifies patients who will not survive, while specificity (true negative rate) reflects how many stable patients are correctly identified as such. Striking the right balance between these two is paramount: a high-sensitivity, moderate-specificity model might be acceptable in a life-or-death scenario, ensuring that most patients at risk receive timely intervention —even if it means some false positives. Conversely, a highly specific model that misses a large portion of at-risk patients may fail to deliver meaningful clinical benefit.

D. Future Efforts

Future efforts should focus on refining model calibration even further, exploring advanced architectures, and integrating new data modalities that may further enhance model sensitivity without the real-world downstream potential of overwhelming clinical staff with false alarms. Threshold tuning and loss function adjustments like Focal Loss can and did help align these predictive models more closely with hypothetical clinical priorities, and further improvements such as prospective validation would benefit an *in actu* deployment of such a system. Ultimately, the goal is to create decision support systems that not only achieve predictive prowess, but also contribute to improving patient outcomes in the intensive care unit setting.

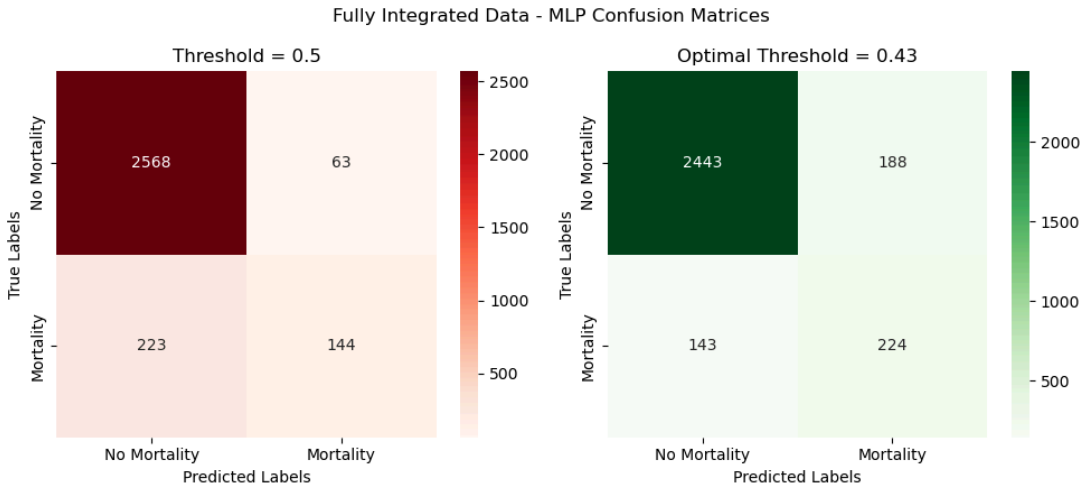

Fig. 3: Comparison of confusion matrices for the default classification threshold (0.5) and the tuned optimal threshold. The left panel uses the default threshold, showing lower sensitivity to the minority class, while the right panel applies the optimized threshold derived from the precision-recall curve, improving minority class detection and overall performance.

TABLE IV: Model Performance Comparison

Integration Level	Model	Minority F1	ROC-AUC
Static Only	Logistic Regression	0.34	0.7573
Static Only	MLP (Thresh. 0.5)	0.04	0.7494
Static + Aggregated Events	Logistic Regression	0.47	0.8516
Static + Aggregated Events	MLP (Thresh. 0.5)	0.07	0.7427
Full Integration	Logistic Regression	0.49	0.8730
Full Integration	MLP (Thresh. 0.5)	0.50	0.8786
Full Integration	MLP (Thresh. Optimized)	0.58	0.8786

References

[1] A. E. W. Johnson *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.

[2] B. L. P. T. H. S. C. L. A. &. M. R. Johnson A., "MIMIC-IV v2.2 — physionet.org," 2023.

[3] D. Li, J. Fu, J. Zhao, J. Qin, and L. Zhang, "A deep learning system for heart failure mortality prediction," *PLOS ONE*, vol. 18, no. 2, p. e276835, Feb. 2023, doi: 10.1371/journal.pone.0276835.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," 2018, [Online]. Available: https://arxiv.org/abs/1708.02002