

حزام السعيدى

Thesis 12 23 2025.docx

- 1- Final Thesis (Word Format) Submission
- Thesis_(2)_AI901_Dr_Hager_Saleh_Third Trimester 2025 (Moodle PP)
- Midocean University

Document Details

Submission ID

trn:oid::1:3450324407

Submission Date

Dec 23, 2025, 10:45 PM GMT+4

Download Date

Dec 24, 2025, 1:19 AM GMT+4

File Name

xcod.663206948_401621_5202_32_21_sisehT_حزام السعيدى_14192

File Size

582.7 KB

28 Pages

6,741 Words

36,846 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





Title

Prepared By

Name: Hizam Ahmed

ID No.: 251002066

Project Title: Development of a multimodal AI framework for ICU outcome mortality prediction using clinical notes and laboratory data

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: Dr. Hager Salah

Collage of Informatics, Midocean University, Comoros

Email: hizam.alsaeedi@gmail.com

Acknowledgements

First and foremost, I thank God for His grace and favor, and I dedicate this research to my late mother, may God have mercy on her. I also remember my father, who passed away during this thesis in the intensive care unit, which only strengthened my resolve to conduct this research so that I could serve patients in intensive care. I also want to express my gratitude to my dear wife and wonderful daughters who stood by me throughout the thesis, supporting and encouraging us, as well as all my brothers and sisters who encouraged and supported us. To them all, I offer my deepest thanks and appreciation. I would also like to express my sincere thanks and gratitude to Dr. Hajar Saleh for her support, guidance, and encouragement throughout the preparation of this research. She always guided me step by step, and thanks to God and then to Dr. Hajar, I overcame all the difficulties and challenges. She played a major role in refining this work and ensuring its successful completion.

I would also like to thank the university and its president for this great institution, which has helped thousands of students graduate and facilitated distance learning. I extend my gratitude to all the faculty members who trained and taught us, enabling us to reach where we are today, and to the colleagues who helped and shared their knowledge during the preparation of this research.

Finally, I offer my sincere thanks to my family and friends for their unwavering love, and to my colleagues at work, including all the administrators and colleagues at Anova for health services and applications, whose support has been a source of strength for me throughout this journey.

Abstract

This research presents a multimodal artificial intelligence framework designed to improve mortality prediction in intensive care units by integrating structured clinical data with unstructured clinical notes. It combines several machine learning techniques with other transformer-based linguistic models to extract numerical patterns and subtle textual information recorded during patient care. For structured data processing, models such as random forests – RF, supporting vector machines - SVM, logistic regression- LR, decision trees - DT, and Naive Bayesian - NV algorithms were employed, utilizing laboratory measurements, vital signs, and demographic characteristics. To handle the unstructured data of the dataset, the thesis used Fine-tuning for each of the BioBERT and ClinicalBERT models, which are well-suited for interpreting the medical language found in clinical observations and have been pre-trained on millions of medical data points, this facilitating and enhancing the research quality. These models transform text into meaningful representations that capture the clinical context. In parallel, the structured variables laboratory results, vital sign are processed separately to generate their own representations.

After encoding both of them, their outputs are merged via a multimodal fusion layer, where the embeddings are combined into a single feature vector. This unified representation is used to predict mortality in the intensive care unit (ICU). The complete framework was trained and tested on a dataset MIMIC-II that included structure data and clinical notes. Its performance was evaluated using several evaluation metrics, including accuracy, precision, recall, F1, Xai and area under the curve (AUC).

The results show that combining Clinical text with structured data features improves predictive performance compared to use either source alone. By combining transformer language models with machine learning techniques, and use self-attention this approach provides a real path toward more reliable mortality prediction. This framework provides a practical foundation for clinical decision support applications and has the potential to improve risk assessment and patient management in ICUs.

Table of Contents

TITLE	1
ACKNOWLEDGEMENTS	2
ABSTRACT	3
LIST OF ABBREVIATIONS	5
LIST OF FIGURE.....	6
LIST OF TABLES.....	6
CHAPTER 1 INTRODUCTION	7
1.1 PROBLEM DEFINITION	7
1.2 THESIS OBJECTIVES	7
1.3 THESIS CONTRIBUTION	8
1.4 THESIS ORGANIZATION	8
CHAPTER 2 OVERVIEW	8
2.1 INTRODUCTION	8
2.2 HEALTHCARE DATA ANALYTICS.....	9
2.3 MULTIMODAL AI IN CLINICAL OUTCOME PREDICTION.....	10
2.3.1 <i>Structured Data Analysis</i>	10
2.3.2 <i>Unstructured Data Analysis</i>	10
2.3.3 <i>Multimodal Fusion Approaches</i>	11
2.3.4 <i>Optimization and Model Tuning</i>	11
CHAPTER 3 RELATED WORK	11
3.1 INTRODUCTION	11
3.2 HISTORY AND LITERATURE REVIEW.....	12
3.3 MULTIMODAL AND FUSION-BASED FRAMEWORKS	12
3.4 OPTIMIZATION METHODS IN MODEL DEVELOPMENT.....	13
3.5 COMPARISON WITH LITERATURE STUDIES.....	13
CHAPTER 4 PROPOSED APPROACH	14
4.1 OVERVIEW	14
4.2 DATA COLLECTION	15
4.3 STRUCTURED DATA	16
4.3.1 <i>Data Preprocessing</i>	16
4.3.2 <i>Training models</i>	16
4.4 UNSTRUCTURED DATA	16
4.4.1 <i>Text Preprocessing</i>	16
4.4.2 <i>Feature Extraction with Transformer Models</i>	17
4.4.3 <i>Fusion model</i>	17
4.5 EVALUATION MODEL	18
4.5.1 <i>Explanation of the ROC Curve</i>	19
4.5.2 <i>Confusion Matrix</i>	19
CHAPTER 5 EXPERIMENTAL RESULTS AND DISCUSSION	19

5.1 EXPERIMENTAL SETUP	19
5.2 RESULTS OF STRUCTURED DATA	20
5.4 RESULTS OF UNSTRUCTURED DATA	20
5.5 RESULTS OF FUSION MODEL	21
5.6 XAI	23
5.7 DISCUSSION.....	24
CHAPTER 6 CONCLUSION AND FUTURE WORK.....	25
6.1 CONCLUSION	25
6.2 FUTURE WORK	26
REFERENCES.....	26

List of Abbreviations

- **Acc.:** Accuracy
- **NLP:** Natural Language Processing
- **ML:** Machine Learning
- **DL:** Deep Learning
- **XAI:** Explainable Artificial Intelligence
- **SVM:** Support Vector Machine
- **RF:** Random Forest
- **DT:** Decision Tree
- **RNN:** Recurrent Neural Network
- **LR:** Logistic Regression
- **AI:** Artificial Intelligence
- **NB:** Naïve Bayes
- **PRE:** Precision
- **REC:** Recall
- **F1:** F1-Score
- **SD:** Structure Data
- **UD:** Unstructured Data

List of Figure

Fig 4. 1 ICU Mortality system architecture	Error! Bookmark not defined.
Fig 5. 1 Accuracy comparison of all models	21
Fig 5. 2 Receiver Operating Characteristic (ROC) curve of the model.....	22
Fig 5. 3 Confusion Matrix.....	23
Fig 5. 4 Xai Emergency Case.....	23
Fig 5. 5 Xai Elective Case.....	24
Fig 5. 6 Accuracy comparison of all models, highlighting the best model	24

List of Tables

<u>Table 3. 1</u> Comparison with literature studies	14
Table 5. 1 Comparison structure models	20
Table 5. 2 Comparison of All Models.....	21
Table 5. 3 Comparison of Multimodal.....	22
<u>Table 5. 4 Comparision Accuracy for models</u>	Error! Bookmark not defined.

Chapter 1 Introduction

1.1 Problem Definition

Predicting patient status in the ICU is one of the most challenges that facing many ICU management in clinical decision-making. Smith et al. (2022) “ICU patients often has complex medical conditions that progress rapidly”, requiring from clinicians to make quick decisions [3]. Traditional methods use a single type of data, often primarily SD such as lab component results and vital signs. The SD provides valuable data but not enough, it does not cover all the important information found in UD clinical records, such as physician notes, progress notes, and discharge summaries [2].

A lot of ICUs departments are not used UD clinical records with SD to followed patient status and marge different data formats. Consequently, current predictive models often lack the capacity to make appropriate decisions and perform poorly in ICU settings [3].

This thesis focused on how to build multimodal AI framework to improve the accuracy of ICU mortality prediction by marge SD and UD, by use ML methods NLP and DL to enhance predictive performance and facilitate the interpretation of clinical outcomes.

1.2 Thesis Objectives

Our objective from this thesis is to build a multimodal AI framework that marge SD and UD for improved ICU mortality prediction.

This objective includes:

- Data preprocessing: we are using both SD and UD datasets to improve data for good training.
- Model development for individual modalities: To develop separate models for SD and UD such as RF, LR, NB, DT, SVM, and RNN for SD, and Word Count, TF-IDF and transformer models for healthcare such as ClinicalBERT for textual data.
- Multimodal fusion: To marge both modalities through a concatenation (fusion) layer to get a comprehensive patient representation for outcome prediction.
- Performance evaluation: we are using standard methods for evaluation such as accuracy, F1-score, recall, precision, and ROC.

1.3 Thesis Contribution

This thesis contributes to the rapidly growing field of healthcare by developing a multimodal model that helps patients visualize their complex medical condition to intensive care physicians. This model integrates SD and UD. Key contributions include:

- Development of a multimodal AI architecture that combines recurrent neural networks (RNNs) for SD processing with a BioBERT/Transformer model to accurately predict a patient's current condition and alert physicians for rapid intervention in emergencies.
- Clean the dataset for preprocessing each of SD and UD that be capable of handling Textual data and structure.
- A fusion model integration strategy for combines deep representations from both data types to improve predictive performance and use self-attention to understand contextual and meaning of data
- A comprehensive performance evaluation achieves higher accuracy, recall, and F1 score compared to traditional single modal.

1.4 Thesis Organization

- **Chapter 2** Overview of ML and data mining in healthcare, focusing on predictive modeling and the role of multimodal data.
- **Chapter 3** reviews related work on ICU outcome prediction, including existing models that use SD, text-based clinical notes, and multimodal fusion techniques.
- **Chapter 4** presents our fusion model detailing for preprocessing, feature extraction, architectures, and the integration between SD and UD.
- **Chapter 5** discusses the experimental setup, implementation details, and comparative evaluation of model performance across multiple baselines.
- **Chapter 6** Thesis Summarizing and expected for future.

Chapter 2 Overview

2.1 Introduction

In recent years, hospitals have been searching for digital systems, especially electronic health records (EHRs), and ai solution to improve quality, this is particularly clear in ICUs, where patient

information is collected almost time and the patient need fast decision [6]. Some of these data points are highly SD lab tests, vital signs, medication dosage while others come in the form of free-text notes written by clinicians [13]. Although these sources look very different, each of them captures something meaningful about a patient's condition.

Working with this type of data is not straightforward. ICU datasets are usually large, messy, and often inconsistent from one patient to the next [5]. Traditional statistical approaches tend to struggle with such complexity, which is one reason why theses have increasingly turned to AI and ML [8]. These methods are better suited for analyzing high-dimensional, multimodal information [16]. In addition, recent progress in DL and NLP has made it possible to extract medical meaning from UD using models that were trained specifically for biomedical language, like some big models BioBERT [7].

The purpose of this chapter is to write a general overview of the concepts that guide the development of multimodal AI models for ICU mortality prediction. I start by describing how different types of healthcare data are organized and used in practice. After that, the chapter discusses the key characteristics of SD and UD ICU data. Finally, I explain why combining these two forms of information can improve prediction performance and provide more useful insights for clinicians.

2.2 Healthcare Data Analytics

Healthcare data analytics is essentially the practice of examining different types of clinical information to understand what is happening with patients and why. In hospitals, especially in intensive care units, a large amount of information is generated every day [5]. Some of it is numerical and well-organized such as laboratory readings, vital signs, or demographic records while other information appears in free-text form, including notes written by doctors, radiology comments, and discharge summaries [16].

There are several steps to data analysis.

1. Collecting and combining data:

The data come from laboratory department and nurse department and EHR, so they must be merged before analysis [3].

2. Preparing the data:

Find missing values the fill it with best solution because we are need all data to be available, then clean text from punctuations [13].

3. Extracting meaningful features:

Select the most feature important that need it for make decision from SD, and in other hand the text will be transfer to numerical representations using language models such as BioBERT, or ClinicalBERT [3].

4. Building predictive models:

Different learning methods ranging from RF and SVM to deep learning approaches like RNNs or Transformer-based models are applied to estimate clinical outcomes [16].

5. Evaluating model:

Accuracy, F1-Score, Recall, Precision, and AUC are commonly used to judge how well a model works and how reliable it is when applied to new patient cases.

In ICU mortality prediction, these steps form the basis of decision-support tools that track patient status and provide early warnings when the risk of deterioration begins to rise [14].

2.3 Multimodal AI in Clinical Outcome Prediction

Early attempts to predict outcomes in the ICU mainly depended on SD, numeric information things like the APACHE II or SOFA scores. These systems were useful, but they only described what could be measured. In recent years, theses have found that models perform noticeably better when they combine those measurable features with the information found in free-text notes written by clinicians [6]. SD reflects state of the patient's physiological in numbers, while UD notes often include reasoning, context, and observations that never appear in a chart of measurements [5].

2.3.1 Structured Data Analysis

SD ICU data includes lab results such as creatinine or hemoglobin and vital signs like heart rate or blood pressure. Traditional ML algorithms DT, RF, SVM, and others have all been tested these features [18].

2.3.2 Unstructured Data Analysis

UD mostly refers to clinical notes. These notes require NLP methods because the information is written in natural language rather than arranged in tables [16]. Models like BioBERT and ClinicalBERT, trained on medical literature and real clinical documents, are able to understand context, terminology, and even the style in which physicians describe their thought process [16]. As a result, they can extract hints about diagnoses, treatments, and day-to-day changes that are difficult to quantify but still crucial for prediction [11].

2.3.3 Multimodal Fusion Approaches

Multimodal learning brings SD and UD data together in a single predictive system. This can be done in different ways: Early fusion mixes all features before training, Intermediate fusion connects representations from separate models for text and numeric data, Late fusion combines outputs from standalone models [16].

Regardless of the strategy, the idea is to form a complete and more realistic picture of the patient. Newer multimodal approaches such as Transformer-Fusion, multimodal transformer models, and hybrids that blend LSTMs with CNNs have shown strong results on commonly used datasets like MIMIC-III and MIMIC-IV [18].

2.3.4 Optimization and Model Tuning

We are using some evaluation methods to be sure about the prediction result from ML algorithm and DL algorithm like accuracy and others methods like ROC and Xai for evaluate fusion model.

Chapter 3 Related Work

3.1 Introduction

The hospitals have a lot of electronic health records. They include lab tests and vital signs and clinical notes. This data helps them to build ai system for improve quality in ICUs mortality

predication and help physician to make decision. The traditional structure not valuable in complex issue so, we need to build multimodal framework to get accurate result.

In this chapter, we are searching for some papers that focus of ICU mortality and we are find Some studies use SD only, while some studies use clinical notes only, and some try to combine both.

3.2 History and Literature Review

ICUs department prediction mostly used SD, like lab results or vital signs. A lot of people see the traditional ML is good for them but he doesn't know when you marge two types of data, the model will be strong with high predication accuracy that will be helpful in physical life [16].

Gao et al. (2024) [2] tested a bunch of ML models DT, Gradient Boosting, SVM, and RF to predict sepsis using MIMIC-IV. The author cleaned the data and handled missing values, then regrouped categories, finally used SMOTE to balance the dataset. RF achieved the best an AUROC of 0.94.

Iwase et al. (2022) [6] “the author compared RF, NN and LR for looking at ICU mortality and length of stay”. The RF achieved the best with AUROC of 0.961. Huddar et al. (2016) [4] did something similar for acute renal failure using MIMIC-II. RF still the best AUC of 0.844.

Some hospital has huge data, so in this case the best solution use DL because it can marge a multi type of data. Al-Dailami et al. (2025) [5] was compared CNNs, RNNs, and attention-based networks using MIMIC-III. They used a Transformer for codes and BERT for clinical notes. CNN achieved the highest AUROC, 0.9149.

Yang et al. (2021) [7] used CNNs with time-series data and clinical text together for disease diagnosis. Their model score was AUCROC of 0.861. The other studies like Shi and Zuo introduced IDDSAM with combining preprocessing strategies for MIMIC-III. Their accuracy was between 55% and 93%.

Kumar et al [9]. The author was reviewed a lot of methods on MIMIC-III. He applied ML methods like regression and SVMs and NN. DT was good in some tasks, but DL was the best for huge data. The authors Huang and Altosaar are did fine-tuned to ClinicalBERT on hospital text. The score was AUROC 0.714.

3.3 Multimodal and Fusion-Based Frameworks

ICU prediction need to marge two type of data to get comprehensive look to all information about this patient.

Ruan et al. (2025) [3] was built a system that can marge two type of data, SD and clinical notes. They used DL algorithm like ResNet and Transformers for the structure data, and fine-tuning models for the text. The model achieved accuracy of 0.7672, AUROC 0.8534, and AUPRC 0.4977.

Caicedo-Torres et al. (2022) [12] The author used CNNs on nursing notes. The ROC-AUC was 0.87.

3.4 Optimization Methods in Model Development

Designing the model is only part of the process choosing the right hyperparameters is equally important. We are using grid search for find the best settings. Adjusting the parameters this way usually improves model performance and helps it generalize more reliably [3]. This step is especially critical for healthcare data, where there are many features, complex interactions, and occasional inconsistencies or noise.

3.5 Comparison with literature studies

The Table 3.1 compare our work with related paper based on models, best model and result using deferent dataset MIMIC-IV, MIMIC II and MIMIC-III to predicate mortality. The author [2] applied LGBM, DT, RF, SVM, XGB and MLP and RF was the best model with 85.99. In [3], the authors applied MLP, ResNet and FT-Transformer and the FT-Trans achieved the high score with 87.32. In [4], the authors applied LR, SVM, DT, AdaBoost and RF and the best model was SVM with 88.58 score. In [11], the authors applied Text-TF-IDF-CNN, LS, DR and TD to achieved 76.33 score. The authors applied [13] Transformer fusion, Clinical BERT, BioBERT and BERT using tow dataset MIMIC-III + eICU-CRD and the Clinical BERT was the best with 85.10. The authors [14] applied CNN, TF-IDF and RNN and the TF-IDF / 1-NN was the best with score 86.27. In [15], the authors applied Self-supervised contrastive + LSTM, LR and Baseline and the best model was Baseline with score 85.5. Our work proposed multimodality framework to improve ICU mortality predication by integrating structure, UD and transformer models. Our model was recorded with 91.13 score.

Paper	Models	Dataset	The best model	Result
[2]	LGBM, DT, RF, SVM, XGB, MLP	MIMIC-IV	RF (Cross Validation)	85.99
[3]	MLP, ResNet, FT-Transformer	MIMIC-IV	FT-Trans	78.32
[4]	LR, SVM, DT, AdaBoost and RF	MIMIC II	SVM	88.58
[11]	CNN, LS, DR, TD	MIMIC-III	CNN	76.33

[13]	Transformer fusion, Clinical BERT, BioBERT , BERT	MIMIC-III + eICU-CRD	Clinical BERT	85.10
[14]	CNN, TF-IDF, RNN	MIMIC-III	TF-IDF / 1-NN	86.27
[15]	Self-supervised contrastive + LSTM, LR, Baseline	MIMIC-III	Baseline	85.5
Our Work	RNN, SVM, DT, RF, NB, LSTM, Transformer, Clinical BERT, BioBERT	MIMIC-III	Fusion	91.13

Table 3. 1 COMPARISON WITH LITERATURE STUDIES

Chapter 4 Proposed Approach

4.1 Overview

This chapter introduces a multimodal AI framework developed to estimate ICU mortality by combining two main sources of patient information: SD such as demographic details and lab results and UD clinical notes.

The key idea behind this framework is that patient outcomes in the ICU depend on both measurable physiological signals and descriptive observations made by medical staff. By bringing these numerical and textual elements together in one model, the system can enhance accurate predictions than approaches that rely on only a single type of data [3].

The overall architecture of the proposed system is illustrated in Figure 4.1.

There are six key stages:

1. Data collection
2. Data Preprocessing for SD and UD
3. Feature representation
4. Modal Training
5. Fusion modal
6. Classification result

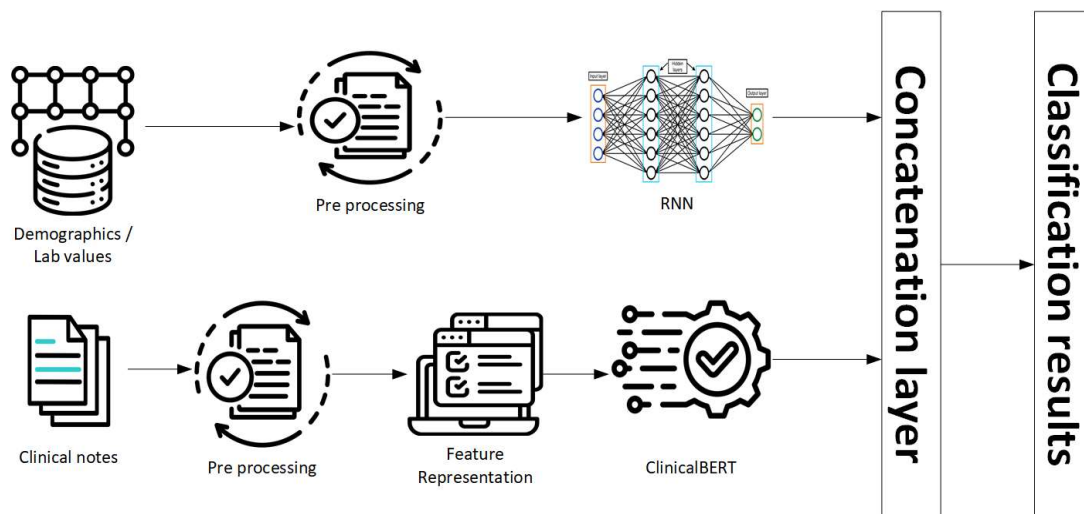


Fig 4. 1 ICU system architecture

4.2 Data Collection

The dataset used in this thesis was download from the MIMIC-III [23]. This dataset contains of two type of data, structure and UD, in additional a lot of medical information specially in ICU so, we are use it to help us for train multimodal with high accuracy.

This dataset contains two type of data. The part one is SD includes demographic details like age, gender, vital sign and laboratory result. For example, some features like heart rate, respiratory rate, blood pressure diastolic, systolic and SpO₂, body T and white blood. The second part is clinical notes written by physicians and nurses.

To define the prediction target, patients were assigned to one of two categories based on the clinical records: 429 cases were labeled as ELECTIVE, while 363 cases were labeled as EMERGENCY and total of samples is 792 samples. Before analysis, several preprocessing steps were carried out. Missing numerical values were replaced using the mean of each feature, while missing value for categorical fields were filled using the best frequent value. Categorical attributes were then encoded to make them suitable for machine-learning models. For the clinical notes, preprocessing involved

tokenization, removal of stop words, case normalization, and elimination of irrelevant symbols and noise and the clinical notes were converted into numerical representations using BioBER [24].

4.3 Structured data

4.3.1 Data Preprocessing

SD are often noisy, incomplete, and heterogeneous in ICU settings. To ensure data quality and consistency, the following preprocessing steps were applied [16]:

1. **Missing Values:** Missing lab values and demographic attributes were imputed using median or mean imputation, depending on variable type.
2. **Encoding Categorical Features:** Non-numeric fields such as gender or label were encoded.
3. **Feature Selection:** reduced variance from features and prevent overfitting (text).

4.3.2 Training models

Once preprocessed, the SD were normalized into numerical data suitable for input into DL models. The ML and DL like:

- **RF:** To capture non-linear feature interactions.
- **DT:** As a baseline interpretable model.
- **SVM:** For high-dimensional boundary-based classification.
- **LR:** As a linear baseline for comparison.
- **NB:** To model probabilistic dependencies.
- **RNN:** Used to learn temporal dependencies and complex patterns.

Each model produced predictive features or class probabilities representing the SD contribution to the overall mortality risk.

4.4 Unstructured data

4.4.1 Text Preprocessing

Clinical notes contain rich contextual information but require careful preprocessing to handle linguistic noise [16].

The following steps were performed:

1. **Tokenization:** We are split text into tokens using a domain tokenizer.

2. **Stopword Removal:** Remove words that not have value in context (e.g., “the,” “and,” “was”) from stopwords list.
3. **Lemmatization:** Return word to the base.
4. **Lowercasing and Cleaning:** Converted text to lowercase and remove punctuation, digits, and special characters were removed.

4.4.2 Feature Extraction with Transformer Models

For advanced text representation, two pre-trained transformer-based biomedical models were used:

- **Word Count [25]:** This method represents clinical notes as a vector of token frequencies. Each feature corresponds to the number of times a specific word appears in the document. Word Count captures basic lexical information and provides a simple, interpretable representation of the text.
- **TF-IDF (Term Frequency–Inverse Document Frequency):** TF-IDF improves upon Word Count by weighing words according to their importance across the corpus. Terms that appear frequently in a particular note but rarely across all notes receive higher weights. This allows the model to emphasize clinically meaningful terms while reducing the impact of common but uninformative words (e.g., “patient,” “the”) [16].
- **BioBERT [24]:** A This model was pre-trained on biomedical literature (PubMed and PMC).
- **ClinicalBERT [24]:** This model has training from hospital discharge summaries and MIMIC-III [23] clinical notes.

Each sentence was tokenized using the corresponding model tokenizer (`AutoTokenizer.from_pretrained`), and the encoded tokens were passed through the transformer layers to obtain contextual embeddings.

The final hidden states from the [CLS] token was extracted as the textual feature vectors representing the patient’s UD.

4.4.3 Fusion model

The fusion model integrates two models:

- **SD:** Patient demographics data vital signs, and lab result.
- **UD (Clinical Notes):** Consists of clinical notes from doctors and nurses describing patient conditions, progress, and diagnoses. After text cleaning and normalization, then applied BioBERT and ClinicalBERT to extract contextual embeddings that capture medical information

This fusion approach is used in multimodal deep learning because it is integration two models RNN and ClinicalBERT into fusion model to encoder data and use self-attention to improve context meaning for data. This integration represents the predication layer for the funal ICU mortality outcome. We are representing our model with this equation fusion = structure + UD.

4.5 Evaluation model

We are evaluating our model with this metrics:

1. Accuracy:
Accuracy measures how well a model is generally correct, and is defined as:
$$(TP + TN) / (TP + TN + FP + FN) [19]$$
2. Precision:
Precision represents the number of positive cases out of all positive cases :
$$TP / (TP + FP) [20]$$
3. Recall (Sensitivity):
Recall represents the number of positive cases divided by total of positive cases and false negative: $TP / (TP + FN) [21]$
4. F1-Score:
The F1-Score is depend of P and R to get the best evaluation for our model :
$$F1 = (2 \times P \times R) / (P + R) [22]$$

These methods were used to measure our model to improve output quality and result.

We are used the activation function like SoftMax and Sigmoid (for binary classification) to output the predicted mortality probability. In the fusion we are using Relu activation function to improve performance:

- **ACC:** General accuracy from all score
- **PRE:** Positive cases
- **REC:** Sensitivity to true positive cases
- **F1:** Balance between P and R
- **AUC:** ROC
- **ROC.**
- **Xai.**

We are using the CV and hyperparameter optimization in grid search for traditional ML models and fine-tuning for DL models to find the optimal solution for each model

4.5.1 Explanation of the ROC Curve

Figure 5.2 presents the ROC curve for our ICU fusion model. This ROC shows the accurate percent to our model. In addition to show the blue line is near of the number one, this mean we process our model with high models and technique to access to final optima solution [26].

- [1] The x-axis represents the False Positive Rate (FPR).
- [1] The y-axis represents the True Positive Rate (TPR).

The ROC shows the blue line still above the diagonal, and this indicating our fusion model is strong for separate output classes. The AUC is 0.94.

- Excellent overall classification performance
- High sensitivity and specificity
- Reliable distinction between ICU and non-ICU cases

These results show that the fusion n is highly effective at correctly identifying ICU outcomes while minimizing misclassifications.

4.5.2 Confusion Matrix

The CM table demonstrates the performance of a model by comparing predicted classifications with actual classifications in a dataset. It presents four key results: TP, TN, FP, and FN, which mean all these methods provide a clear picture of the types of errors in the model. Because it highlights both correct decisions and incorrect classifications, this table is very important for predication evaluation.

Chapter 5 Experimental Results and Discussion

This chapter presents our result and experimental setup for our model. The analysis was performed on a clinical dataset comprising two complementary data modalities: SD, and UD textual data. To assess the independent contribution of each data modality, a range of ML and DL models were trained separately using the SD variables and the textual features. Following the unimodal evaluations, feature representations from both modalities were integrated into a unified multimodal learning architecture. Model performance was systematically evaluated using classification metrics, including Accuracy, F1-score, Recall, Precision, and AUC.

5.1 Experimental Setup

The model were done on Google Colab using a GPU, using PyTorch and scikit-learn. The data that used for training 80 and for testing 20. For the classical models, the parameters of ML models were optimized by GridSearchCV. For DL models, Batch size 32, the number of Epochs is 40 with early

stopping, loss function, AdamW optimizer. The data was split into 80%for training 634 samples and 20%for testing 158 samples.

5.2 Results of structured data

Table 5.1 shows the deferent ML RF, SVM, DT, LR and NB with deferent accuracy precision recall and F1-score. We notice the LR get the best accuracy 71.85 and best precision 72.25 and best recall 71.85 and best F1-score 71.90. The RF model followed closely behind with a performance level just below LR. Its accuracy of 70.59%, The SVM model performed at a moderate level, reaching 67.23% accuracy. On the other hand, the DT result was the lower score 62.61% accuracy. Finally, NB achieved 63.87% accuracy and the lowest F-score among the models.

Models	Acc	P	R	F-score
RF	70.59	71.05	70.59	70.64
SVM	67.23	67.15	67.23	67.17
DT	62.61	62.67	62.61	62.65
LR	71.85	72.25	71.85	71.90
NB	63.87	64.40	63.87	62.37

Table 5. 1 Comparison structure models

5.4 Results of unstructured data

Table 5.2 shows the results of ML models using two feature representation: word count and TF-IDF, using different evaluation methods: accuracy Using word count features, LR recorded the best accuracy at 90.57% ac, with SVM close behind at 89.31%. Both models had balanced Precision, Recall, and F1-scores, showing that even basic text features can be useful when combined with solid linear models. On the other hand, DT, RF, and NB scored lower (around 84–85%), which suggests that word counts might not fully capture the context needed for more complicated clinical notes. We can see the second-best model is SVM in word count recorded the second performance. Also, we can see the RF and the DT reach the same similarity because to include the trees. LR and SVM in TF-IDF were the best model and record the same result of the Acc, P, R and F-score. The better performance of ClinicalBERT was achieved good result comparing with BioBERT and The Precision was the highest score for all other models.

Feature representation	Models	Accuracy	Precision	Recall	F1-score
	RF	84.28	84.27	84.28	84.27

Word count	SVM	89.31	89.33	89.31	89.29
	DT	84.91	84.96	84.91	84.92
	LR	90.57	90.56	90.57	90.56
	NB	85.53	85.53	85.53	85.53
TF-IDF	RF	85.53	85.53	85.53	85.53
	SVM	89.94	89.94	89.94	89.94
	DT	79.25	79.23	79.25	79.23
	LR	89.94	89.94	89.94	89.94
	NB	84.28	84.3	84.28	84.28
BERT	Bio BERT	84.08	91.23	72.22	80.62
	Clinical BERT	85.35	98.04	69.44	81.30

Table 5. 2 Comparison of All Models

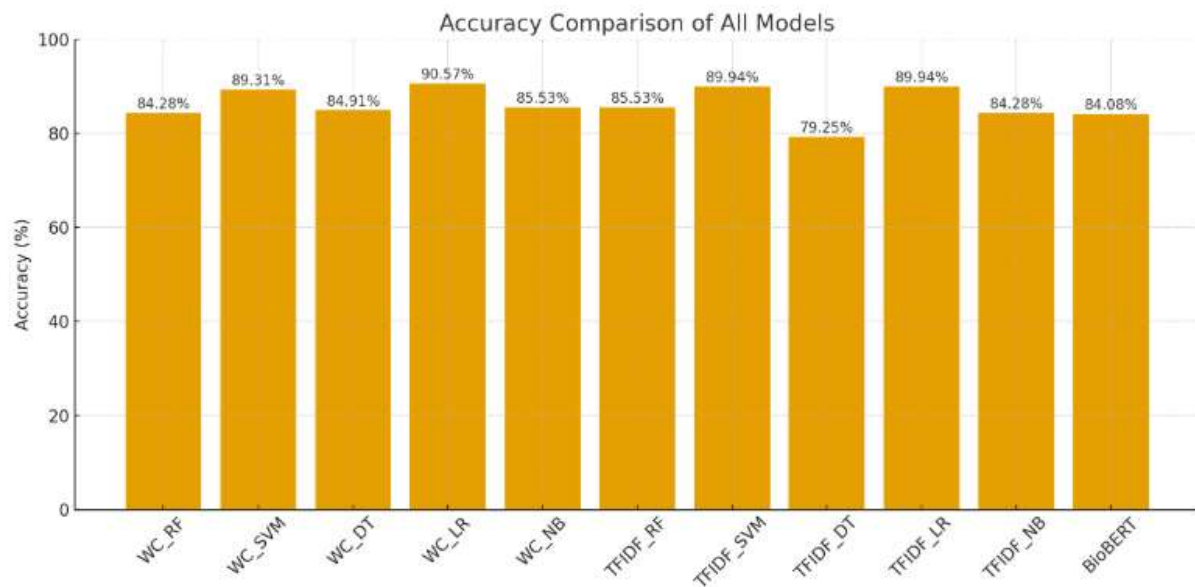


Fig 5. 1 Accuracy comparison of all models

5.5 Results of fusion model

The Fusion model achieved the best score comparing with all other models, the result was 91.08% accuracy, 91.13% precision, 91.08% recall, and an F1-score of 91.09%. The balance for all methods

gives us clear image the model was training correctly. We can capture complex relation between fin tuning model and other models by transformer design.

Feature representation	Accuracy	Precision	Recall	F1-score
RNN	58.40	69.99	58.40	47.94
Clinical BERT	85.35	98.04	69.44	81.30
Multimodal	91.08	91.13	91.08	91.09

Table 5. 3 Comparison of Multimodal

In the figure 5.2, [26] Fawcett, T. (2006) “the blue curve illustrates the model’s actual performance, while the red dashed diagonal line represents a classifier with no discriminative ability” (essentially random guessing), where the AUC would equal 0.50. An AUC score of 0.94 suggests that the model correctly ranks a randomly chosen positive instance higher than a negative one 94% of the time, which is considered very good performance in clinical prediction tasks.

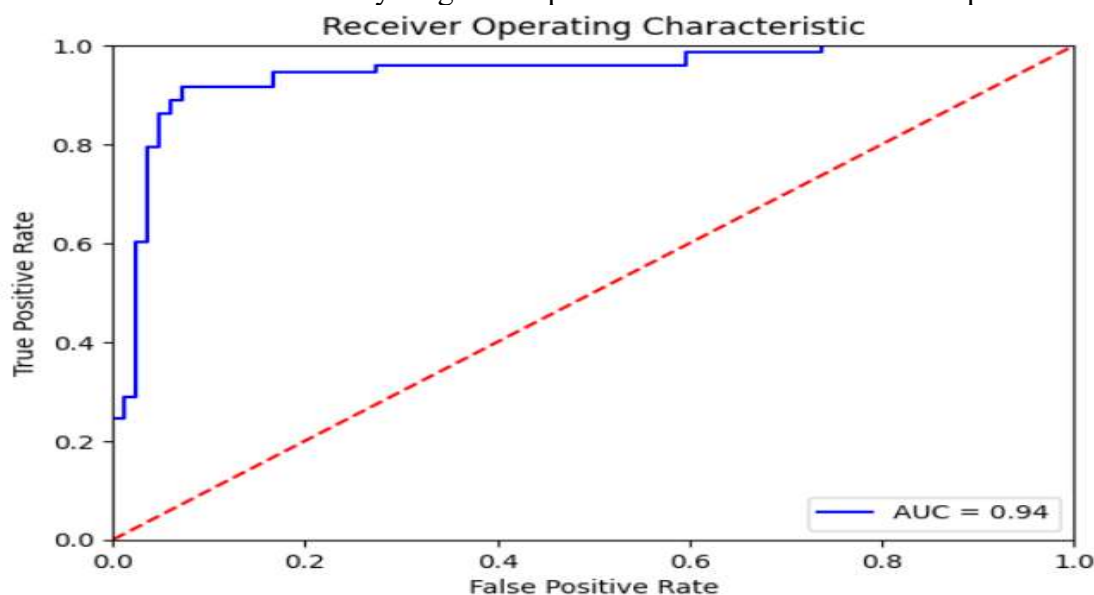


Fig 5. 2 Receiver Operating Characteristic (ROC) curve of the model

In the figure 5.3, The CM shows that the model is particularly good at capture emergency cases, with a Recall of 94.05%. In a medical setting, this is crucial because it keeps the number of missed emergencies—false negatives—very low at about 5.95%. The downside is a higher false-positive rate of 16.44%, meaning some non-emergency cases are mistakenly classified as urgent. Depending on the clinical priority, especially when safety is the main concern, this balance may be acceptable. The model tends to avoiding dangerous misses, even if it results in a few extra cases being flagged unnecessarily.

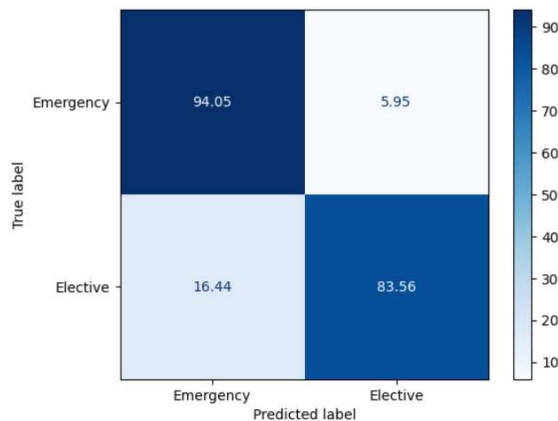


Fig 5. 3 Confusion Matrix

5.6 XAI

We were used two cases EMERGENCY and ELECTIVE to evaluate output for our training model. We were achieved the impressive impact for output highlighted and predication probabilities. In EMERGENCY class was 84% score with right predicate and 16% was ELECTIVE probability. In the ELECTIVE case we achieved 79% for right probability and 21% was EMERGENCY.

- **EMERGENCY Case**

Figure 5.4 shows an example in which the free-text notes clearly indicate that the patient is in a critical condition. The model classified the case as EMERGENCY with a probability of 0.84, mainly because the narrative contains phrases linked to severe injury and the need for rapid action. Words such as “patient,” “is,” and “immediate” carried strong positive influence toward the emergency outcome. This example illustrates how the model can highlight and rely on key clinical expressions in the notes even when some of those words are very common because their meaning

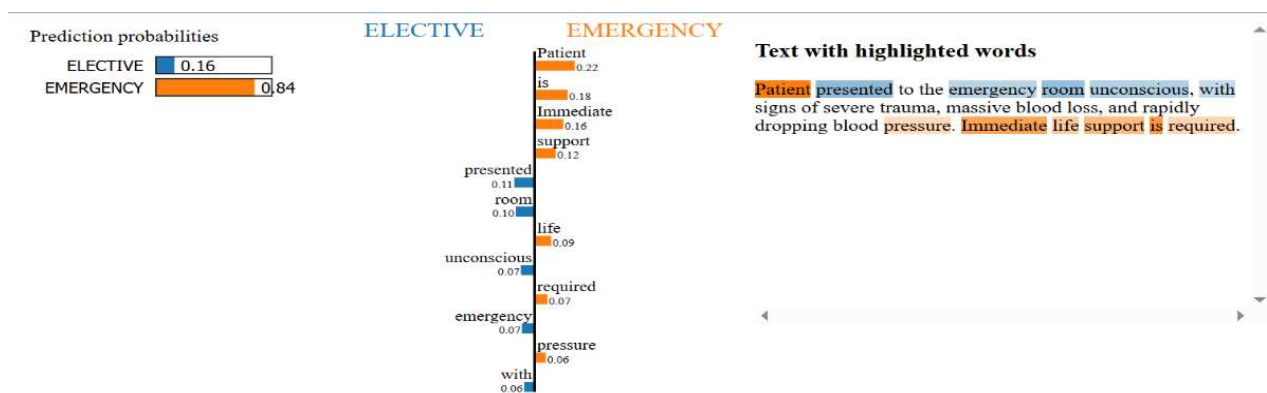


Fig 5. 4 Xai Emergency Case

- **ELECTIVE case** Figure 5.5 shows a word-by-word explanation for one specific prediction. Even though the clinical note includes several expressions that normally point to an acute

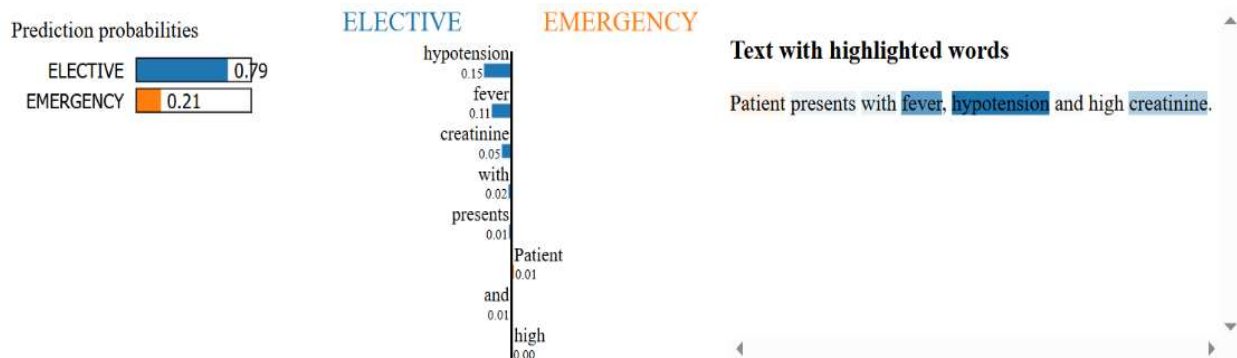


Fig 5. 5 Xai Elective Case

condition such as fever, hypertension, and raised creatinine and these terms lean the model toward an EMERGENCY outcome (with “hypertension” adding about 0.15), the final decision is still ELECTIVE with a probability of 0.79. This means that information from the SD, or other parts of the note that carry weaker influence, outweighed those warning terms.

5.7 Discussion

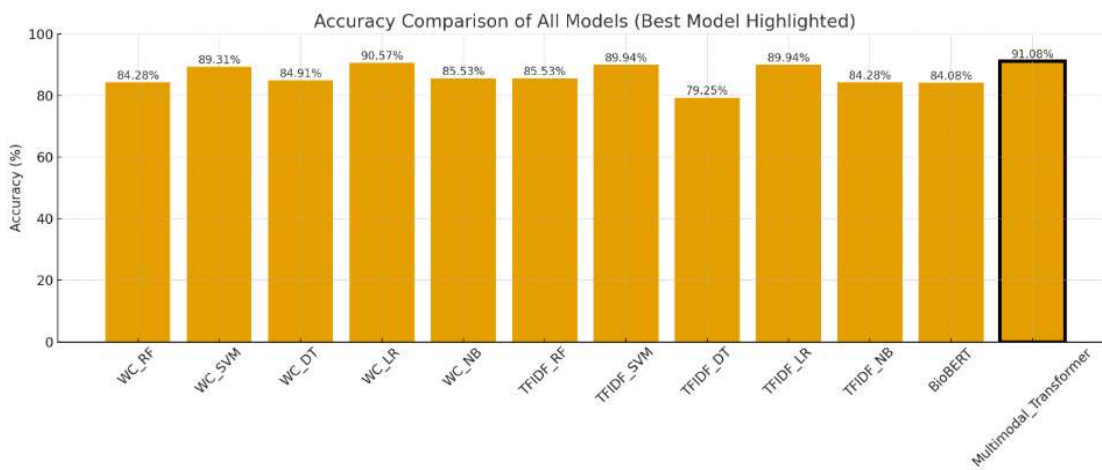


Fig 5. 6 Accuracy comparison of all models, highlighting the best model

Overall, the results showed that combining multiple data modalities improves prediction beyond what individual data sources can achieve. ClinicalBERT provides a strong contextual

understanding of clinical text. The model captures all medical meaning for patient. Multimodal achieve the high accurate for predication and get the best result for improve output predication.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This research improves the accuracy of ICU mortality prediction by integrating SD and UD from patient history. Single data type is often insufficient for accurate results. Therefore, we employed a multimodal framework model combining two sources: structure data such as lab results and vital signs, and clinical notes. The BioBERT and ClinicalBERT models were used to extract meaning from clinical notes, in addition DL models handled SD. We are using ML method for analysis the structure data and get the different result for some algorithm.

The word representation models like word count and TF-IDF to understand every word in the context clinical notes. After training of different models, we believe the multimodal was optimal solution for understand context of text and structure data with result precision 91.13, accuracy 91.08, recall 91.08, F1-score 91.09, and AUC 95. These results highlight the importance of using specialized transformer models to understand clinical notes and integrating them with structure data. In conclusion, this thesis demonstrates that multimedia learning has proven its predictive power in critical care settings and significantly improves prediction accuracy.

6.2 Future Work

Future work can focus on

- 1- Expand the dataset.
- 2- Using Time series
- 3- Test this model in multiple hospital.

References

- [1] Terwiesch, C., KC, D., Kahn, J.M.: Working with capacity limitations: operations management in critical care. *Critical Care* 15, 1–6 (2011)
- [2] Gao, J., Lu, Y., Ashrafi, N., Domingo, I., Alaei, K., Pishgar, M.: Prediction of sepsis mortality in icu patients using machine learning methods. *BMC Medical Informatics and Decision Making* 24(1), 228 (2024)
- [3] Ruan, Y., Tan, D. J., Ng, S. K., Huang, L., & Feng, M. (2025). Towards accurate and reliable ICU outcome prediction: a multimodal learning framework based on belief function theory using structured EHRs and free-text notes. <http://arxiv.org/abs/2501.04389>
- [4] V. Huddar, B. K. Desiraju, V. Rajan, S. Bhattacharya, S. Roy and C. K. Reddy, "Predicting Complications in Critical Care Using Heterogeneous Clinical Data," in *IEEE Access*, vol. 4, pp. 7988-8001, 2016, doi: 10.1109/ACCESS.2016.2618775.
- [5] A. Al-Dailami, H. Kuang and J. Wang, "Multimodal Representation Learning Based on Personalized Graph-Based Fusion for Mortality Prediction Using Electronic Medical Records," in *Big Data Mining and Analytics*, vol. 8, no. 4, pp. 933-950, August 2025, doi: 10.26599/BDMA.2024.9020099.
- [6] Iwase, S., Nakada, T. aki, Shimada, T., Oami, T., Shimazui, T., Takahashi, N., Yamabe, J.,

Yamao, Y., & Kawakami, E. (2022). Prediction algorithm for ICU mortality and length of stay using machine learning. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-022-17091-5>

[7] Yang, H., Kuang, L., & Xia, F. Q. (2021). Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1). <https://doi.org/10.1186/s13326-021-00235-3>

[8] Shi, Z., Zuo, W., Liang, S., Zuo, X., Yue, L., & Li, X. (2020). IDDSaM: an Integrated Disease Diagnosis and Severity assessment Model for Intensive Care Units. *IEEE Access*, 8, 15423–15435. <https://doi.org/10.1109/aCCESS.2020.2967417>

[9] Kumar, S., Oh, I., Schindler, S., Lai, A. M., Payne, P. R. O., & Gupta, A. (2021). Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: A systematic literature review. In *JAMIA Open* (Vol. 4, Issue 3). Oxford University Press. <https://doi.org/10.1093/jamiaopen/ooab052>

[10] Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. <http://arxiv.org/abs/1904.05342>

[11] Xu, K., Lam, M., Pang, J., Gao, X., Band, C., Mathur, P., Papay, F., Khanna, A. K., Cywinski, J. B., Maheshwari, K., Xie, P., & Xing, E. P. (2019). Multimodal Machine Learning for Automated ICD Coding. In *Proceedings of Machine Learning Thesis* (Vol. 106). <https://www.cdc.gov/nchs/icd/icd10cm.htm>

[12] Caicedo-Torres, W., & Gutierrez, J. (2022). ISeeU2: Visually interpretable mortality prediction inside the ICU using deep learning and free-text medical notes. *Expert Systems with Applications*, 202, 117190.

[13] Lyu W, Dong X, Wong R, Zheng S, Abell-Hart K, Wang F, Chen C. A Multimodal Transformer: Fusing Clinical Notes with Structured EHR Data for Interpretable In-Hospital Mortality Prediction. *AMIA Annu Symp Proc*. 2023 Apr 29;2022:719-728. PMID: 37128451; PMCID: PMC10148371.

[14] Shukla, S. N., & Marlin, B. M. (2020). Integrating physiological time series and clinical notes with deep learning for improved ICU mortality prediction. *arXiv preprint arXiv:2003.11059*.

[15] King, R., Yang, T., & Mortazavi, B. J. (2023, December). Multimodal pretraining of medical time series and notes. In *Machine Learning for Health (ML4H)* (pp. 244-255). PMLR.

[16] Saleh, H., Mccann, M., El-Sappagh, S., & Breslin, J. G. (n.d.). TransformerFusionNet: A Real-Time Multimodal Framework for ICU Heart Failure Mortality Prediction Using Big Data Streaming.

- [17] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [18] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.
- [19] S. Liu, X. Wang, Y. Hou, G. Li, H. Wang, H. Xu, Y. Xiang, and B. Tang, "Multimodal data matters: language model pre-training over structured and unstructured electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 504–514, 2022.
- [20] A. M. Merchant, N. Shenoy, S. Lanka, and S. Kamath, "Ensemble neural models for icd code prediction using unstructured and structured healthcare data," *Heliyon*, vol. 10, no. 17, 2024.
- [21] A. Spark, "Apache spark," <https://spark.apache.org/>, online; accessed 2024.
- [22] A. Kafka, "Apache kafka," <https://kafka.apache.org/>, online; accessed 2024.
- [23] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [25] K. P. Chodey and G. Hu, "Clinical text analysis using machine learning methods," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 2016, pp. 1–6, doi: 10.1109/ICIS.2016.7550908.
- [26] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861–874.