

حزام السعيدى

Last Theses.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:333464540

Submission Date

Dec 4, 2025, 11:10 PM GMT+2

Download Date

Dec 15, 2025, 6:04 PM GMT+2

File Name

Last Theses.docx

File Size

601.3 KB

31 Pages**7,692 Words****44,918 Characters**

58% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

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.

Detection Groups

-  **65 AI-generated only 58%**
Likely AI-generated text from a large-language model.
-  **0 AI-generated text that was AI-paraphrased 0%**
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

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 would like to express my deepest gratitude to Dr. Hager Salah for her continuous support, valuable guidance, and encouragement throughout the course of this thesis. Her expertise, insightful feedback, and patience have been instrumental in shaping this work and ensuring its successful completion.

I would also like to thank all the faculty members and colleagues who provided assistance and shared their knowledge during the development of this thesis. Their constructive suggestions and motivation have greatly contributed to the improvement of this study.

Finally, I am sincerely thankful to my family and friends for their constant love, understanding, and support, which have been a source of strength throughout this journey.

Abstract

This these presents a multimodal artificial intelligence framework designed to improve ICU mortality prediction by integrating structured clinical data with unstructured clinical notes. The proposed approach combines multiple machine-learning methods with transformer-based language models to capture both numerical patterns and nuanced textual information recorded during patient care. In the structured data component, models such as Random Forest, Support Vector Machines, Logistic Regression, Decision Trees, and Naïve Bayes are evaluated using laboratory measurements, vital signs, and demographic attributes. To handle the unstructured part of the dataset, the study makes use of BioBERT and ClinicalBERT, which are well suited for interpreting medical language found in clinical notes. These models convert the narrative text into meaningful representations that capture clinical context and subtle linguistic cues. In parallel, the structured variables—laboratory findings, physiological measurements, and demographic attributes—are processed separately to form their own feature representation.

Once both streams are encoded, their outputs are brought together through a multimodal fusion layer that simply joins the two embeddings into one combined feature vector. This unified representation is then used to predict ICU mortality. The full framework is trained and tested on a dataset that includes both numeric patient information and written clinical documentation. Its performance is examined through several evaluation measures, including accuracy, precision, recall, F1-score, and AUC.

The results consistently show that merging text-based embeddings with structured clinical features leads to stronger predictive performance than relying on either source alone. By pairing transformer-based language models with established machine-learning methods, the approach offers a realistic path toward more dependable mortality prediction. The framework provides a practical foundation for clinical decision-support applications and has the potential to enhance risk assessment and patient management within intensive-care settings.

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	9
2.1 INTRODUCTION	9
2.2 HEALTHCARE DATA ANALYTICS AND MINING	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>	11
2.3.3 <i>Multimodal Fusion Approaches</i>	11
2.3.4 <i>Optimization and Model Tuning</i>	11
CHAPTER 3 RELATED WORK	12
3.1 INTRODUCTION	12
3.2 HISTORY AND LITERATURE REVIEW.....	12
3.3 MULTIMODAL AND FUSION-BASED FRAMEWORKS	13
3.4 OPTIMIZATION METHODS IN MODEL DEVELOPMENT.....	14
3.5 COMPARISON WITH LITERATURE STUDIES.....	14
CHAPTER 4 PROPOSED APPROACH	15
4.1 OVERVIEW	15
4.2 DATA COLLECTION	16
4.3 STRUCTURED DATA	16
4.3.1 <i>Data Preprocessing</i>	16
4.3.2 <i>Training models</i>	17
4.4 UNSTRUCTURED DATA	17
4.4.1 <i>Text Preprocessing</i>	17
4.4.2 <i>Feature Extraction with Transformer Models</i>	17
4.4.3 <i>Fusion model</i>	18
4.5 EVALUATION MODEL	19
4.5.1 <i>Explanation of the ROC Curve</i>	19
4.5.2 <i>Confusion Matrix</i>	20
CHAPTER 5 EXPERIMENTAL RESULTS AND DISCUSSION	21

5.1 EXPERIMENTAL SETUP	21
5.2 RESULTS OF STRUCTURED DATA	21
5.4 RESULTS OF UNSTRUCTURED DATA	22
5.5 RESULTS OF FUSION MODEL	23
5.6 XAI	25
5.7 DISCUSSION.....	26
CHAPTER 6 CONCLUSION AND FUTURE WORK.....	28
6.1 CONCLUSION	28
6.2 FUTURE WORK	28
REFERENCES.....	29

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

List of Figure

Fig 4. 1 Proposed system architecture	Error! Bookmark not defined.
Fig 5. 1 Accuracy comparison of all models	23
Fig 5. 2 Receiver Operating Characteristic (ROC) curve of the model.....	24
Fig 5. 3 Confusion Matrix.....	25
Fig 5. 4 Xai Emergency Case.....	25
Fig 5. 5 Xai Elective Case.....	26
Fig 5. 6 Accuracy comparison of all models, highlighting the best model	26

List of Tables

Table 3. 1 Comparison with literature studies	14
Table 5. 1 Comparison structure models	Error! Bookmark not defined.
Table 5. 2 Comparison of All Models.....	23
Table 5. 3 Comparison of Multimodal.....	24
Table 5. 4 Accuracy comparison of all models, highlighting the best model ..	Error! Bookmark not defined.

Chapter 1 Introduction

1.1 Problem Definition

Predicting patient outcomes in the Intensive Care Unit (ICU) is one of the most critical challenges in clinical decision support. ICU patients often have complex medical conditions that evolve rapidly, requiring physicians to make time-sensitive and data-driven decisions [16]. Traditionally, outcome prediction models rely primarily on structured data such as demographics, laboratory test results, and vital signs. Although these structured data elements provide valuable quantitative insights, they fail to capture the qualitative and contextual information found in unstructured clinical notes, such as physicians' observations, progress reports, and discharge summaries [2].

The integration of unstructured clinical text with structured data remains a significant challenge due to differences in data formats, semantics, and dimensionality. As a result, existing predictive models often exhibit limited generalization and lower performance in real-world ICU settings [16].

Therefore, the central thesis problem of this study is: How can structured (laboratory and demographic) data and unstructured (clinical notes) data be effectively integrated into a unified multimodal AI framework to improve ICU mortality outcome prediction accuracy and reliability?

This thesis seeks to bridge the gap between numeric and textual information, taking advantage of recent developments in natural language processing (NLP) and deep learning to enhance predictive performance and clinical interpretability.

1.2 Thesis Objectives

The main objective of this thesis is to design and develop a multimodal AI framework that integrates structured and unstructured clinical data for improved ICU mortality prediction.

The specific objectives include:

1. **Data preprocessing and representation:** To preprocess and clean both structured (lab values, demographic data) and unstructured (clinical notes) datasets to ensure consistency, accuracy, and readiness for modeling.
2. **Model development for individual modalities:** To develop separate models for structured and unstructured data such as Random Forests (RF), Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM), and Recurrent Neural Networks (RNN) for structured data, and Word Count, TF-IDF and transformer-based models such as BioBERT and ClinicalBERT for textual data.

3. **Multimodal fusion:** To integrate outputs from both modalities through a concatenation (fusion) layer to form a comprehensive patient representation for outcome prediction.
4. **Performance evaluation:** To evaluate the proposed model using standard metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).
5. **Comparative analysis:** To compare the multimodal model against unimodal (single-type data) baselines to demonstrate its superiority in ICU mortality prediction tasks.

1.3 Thesis Contribution

This thesis contributes to the growing field of AI-driven healthcare analytics by introducing an innovative multimodal approach that leverages both structured and unstructured ICU data. The key contributions include:

1. **Developing multimodal AI architecture** combining RNN for structured data processing and BioBERT/Transformer for textual feature extraction from clinical notes.
2. **Applying data preprocessing pipeline** capable of handling heterogeneous ICU data, including missing values, irregular sampling, and unstructured text.
3. **A fusion-based integration strategy** using a concatenation layer that effectively combines deep representations from both modalities for enhanced predictive performance.
4. **Comprehensive performance evaluation**, showing that the proposed model achieves higher accuracy, precision, recall, and F1-score compared to traditional unimodal models.
5. **Clinical interpretability**, highlighting how multimodal fusion helps capture both quantitative and narrative aspects of patient health, supporting more informed ICU decision-making.

1.4 Thesis Organization

The remainder of this thesis is organized as follows:

- **Chapter 2** provides an overview of machine learning 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 structured data, text-based clinical notes, and multimodal fusion techniques.
- **Chapter 4** presents the proposed multimodal AI framework, detailing the data preprocessing, feature extraction, model architectures, and the fusion mechanism used to integrate structured and unstructured information.
- **Chapter 5** discusses the experimental setup, implementation details, and comparative evaluation of model performance across multiple baselines.
- **Chapter 6** concludes the thesis by summarizing key findings, highlighting limitations, and outlining future directions for improving multimodal healthcare prediction systems.

Chapter 2 Overview

2.1 Introduction

In recent years, hospitals have been relying more heavily on digital systems, especially electronic health records (EHRs), and this shift has significantly changed daily clinical practice [3]. This is particularly clear in Intensive Care Units (ICUs), where patient information is collected almost continuously [6]. As a result, a single ICU stay can generate an enormous amount of data [13]. Some of these data points are highly structured 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 artificial intelligence (AI) and machine learning (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 clinical meaning from unstructured text using models that were trained specifically for biomedical language, such as BioBERT and ClinicalBERT [7].

The purpose of this chapter is to give 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 structured and unstructured 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 and Mining

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]. Bringing these different elements together is an important part of understanding a patient's overall condition.

When hospitals perform data mining, the work usually moves through several steps.

1. Collecting and combining data:

Information is gathered from electronic health records, lab systems, and bedside monitoring equipment. Often these sources store data differently, so they must be merged before analysis [16].

2. Preparing the data:

At this stage, missing values are handled, abnormal entries are checked, and numerical readings are standardized. Text is also cleaned and converted into a form that can be processed by computational models [13].

3. Extracting meaningful features:

Important variables are selected from structured data, while text is turned into numerical representations using language models such as BERT, BioBERT, or ClinicalBERT [16].

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 performance:

Accuracy, Precision, Recall, F1-Score, 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 structured, 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 [16]. Structured data reflects the patient's physiological state in numbers, while unstructured notes often include reasoning, context, and observations that never appear in a chart of measurements [16]. Bringing these two sources together has become one of the main directions in improving ICU prediction [16].

2.3.1 Structured Data Analysis

Structured ICU data includes lab results such as creatinine or hemoglobin, vital signs like heart rate or blood pressure, and demographic factors. Traditional machine-learning algorithms DT, RF,

SVM, and others have all been widely tested on these features [18]. These models are especially good at picking up patterns and relationships that are not obvious from looking at the raw numbers alone, which helps in estimating a patient's survival risk [18].

2.3.2 Unstructured Data Analysis

Unstructured data 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 structured and unstructured 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

To improve performance, these usually fine-tune hyperparameters through methods like grid search, Bayesian optimization, or SHAC. These techniques help identify model settings that maximize metrics such as AUC or F1-Score, ensuring that the model performs well not only on the training data but also on new ICU cases [16].

Chapter 3 Related Work

3.1 Introduction

These days, hospitals have a lot of electronic health records. They include numbers, like lab tests and vital signs, but also doctors' and nurses' notes. Having both makes it possible to build AI systems that try to predict what might happen to patients in the ICU. Before, people mostly used simple statistical models. The basic methods worked to some extent, but they struggled with the more complicated patterns in the dataset. Machine learning and deep learning models do a better job, since they consider both the numbers and the text and can uncover hidden connections.

In this chapter, we go over recent studies on ICU predictions, like mortality or sepsis. Some models only use structured data, some only use clinical notes, and some try to combine both. These also spend a lot of time tweaking the models, merging different sources of information, and using newer architectures like transformers. All of this helps make predictions more reliable and closer to what doctors actually see.

3.2 History and Literature Review

In the beginning, ICU prediction mostly looked at structured data, like lab results or vital signs. People used traditional ML methods LR, DT, SVM. They worked okay, but as computers got faster and more data became available, DL and multimodal models started showing up. These could handle structured data and clinical notes at the same time [16].

Some studies stand out. Gao et al. (2024) [2] tested a bunch of machine learning models DT, Gradient Boosting, XGBoost, LightGBM, Multilayer Perceptrons, SVM, and RF to predict sepsis using MIMIC-IV. They cleaned the data, handled missing values, regrouped categories, and even used SMOTE to balance the dataset. RF came out on top with an AUROC of 0.94.

Iwase et al. (2022) [6] also compared RF, XGBoost, Neural Networks, and LR, looking at ICU mortality and length of stay. Again RF did best AUROC of 0.961 for short stays, 0.830 for long ones. Huddar et al. (2016) [4] did something similar for acute renal failure using MIMIC-II. Despite messy data, RF still outperformed others, with an AUC of 0.844.

When EHR data became more complex, DL gained ground. Al-Dailami et al. (2025) [5] compared CNNs, RNNs, autoencoders, and attention-based networks for ICU outcomes using MIMIC-III. They used a Transformer-based skip-gram for codes and BERT for clinical notes. CNN ended up with the highest AUROC, 0.9149.

Yang et al. (2021) [7] used CNNs to look at time-series data and clinical text together for disease progression. Their model got an AUCROC of 0.861. Shi and Zuo introduced IDDSAM, combining

preprocessing strategies for MIMIC-III. Their accuracies ranged from 55% to 93%, doing well in diagnosis and mortality prediction.

Kumar et al [9]. reviewed lots of methods on MIMIC-III from regression and SVMs to Bayesian and neural networks. DT did well in some tasks, but deep learning scaled better with bigger datasets. Huang and Altosaar fine-tuned ClinicalBERT on hospital text. Unlike regular BERT, it caught medical nuances, with AUROC 0.714, and gave interpretable results through attention weights.

3.3 Multimodal and Fusion-Based Frameworks

Lately, ICU prediction has been moving toward combining numbers and notes to get a fuller picture of patients.

Ruan et al. (2025) [3] built a system that merged structured data with clinical notes. They used ResNet and Transformers for the numbers, and pre-trained language models for the text. The model got balanced accuracy of 0.7672, AUROC 0.8534, and AUPRC 0.4977.

Xu et al. [11] made an ensemble model. Text-CNN handled the notes, deep learning took care of semi-structured data, and decision trees worked on the tables. On MIMIC-III, it reached a micro-AUC of 0.9541, beating most of the baseline models.

Caicedo-Torres et al. (2022) [12] created ISeeU. It used multi-scale CNNs on nursing notes to predict ICU mortality. They added Shapley values to explain the predictions. ROC-AUC was 0.8735. Clinicians could actually understand what the model was doing.

Lyu W et al. (2022) [13] used a Multimodal Transformer. It combined structured EHR data and notes, capturing both time patterns and context. On MIMIC-III and eICU-CRD, AUROC was 0.877, AUPRC 0.538. Better than other fusion methods.

Shukla et al. (2019) [14] combined LSTM for time-series data with CNN for notes. AUROC reached 0.90. Shows mixing sequential and text data works well.

King et al. (2023) [15] went further with self-supervised pretraining. They used contrastive learning and masked modeling to encode both physiological and text data. Needed very little labeled data and improved AUC-ROC by 0.17 over models trained from scratch.

Finally, Saleh et al. (2023) [16] made a real-time system for ICU heart disease prediction. Transformer-Fusion combined BioBERT for notes and RNNs for numbers. They used MIMIC-III. Accuracy reached 91.72%. Shows Transformer-based models are really powerful in clinical AI.

3.4 Optimization Methods in Model Development

Designing the model is only part of the process choosing the right hyperparameters is equally important. Methods such as grid search, Bayesian optimization, and SHAC are commonly used to find the best settings. Adjusting the parameters this way usually improves model performance and helps it generalize more reliably [16]. 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, unstructured data 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]	Text-TF-IDF-CNN, LS, DR, TD	MIMIC-III	Text-TF-IDF-CNN + LS + DR + TD	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: structured data such as demographic details and laboratory results and unstructured 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 offer more accurate predictions than approaches that rely on only a single type of data [16].

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

It consists of six key stages:

1. Data collection
2. Preprocessing of structured and unstructured data
3. Feature representation and model design
4. Individual model training for each modality
5. Multimodal fusion via concatenation
6. Final classification and evaluation

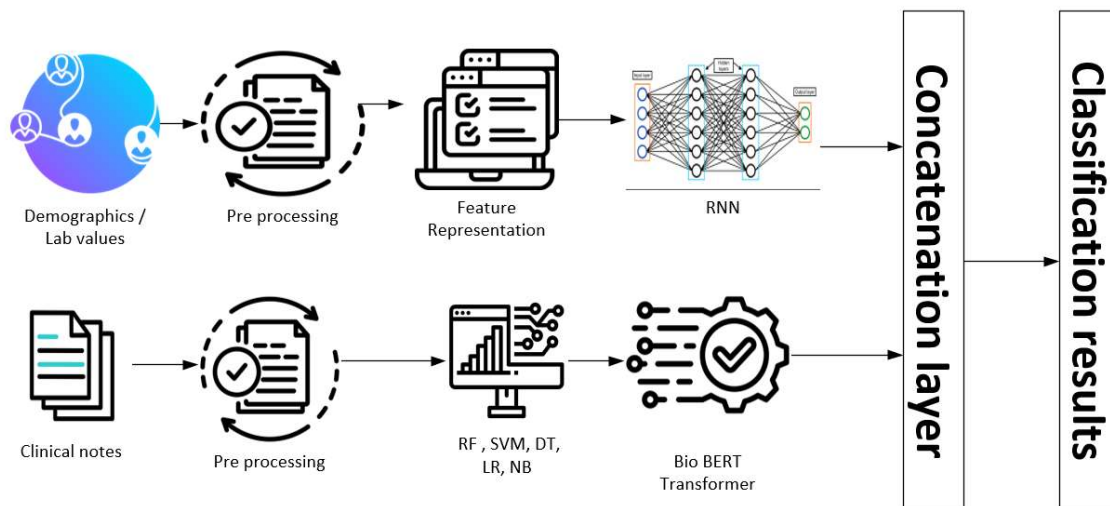


Fig 4. 1 Proposed system architecture

4.2 Data Collection

The dataset used in this study was obtained from the MIMIC-III [23] clinical database, which provides de-identified information for patients admitted to the intensive care unit. For this work, both structured records and narrative clinical notes were collected to build a multimodal view of each patient's condition.

The structured part of the dataset includes demographic details such as age, and gender, along with a wide range of physiological and laboratory measurements. These measurements cover heart rate, systolic and diastolic blood pressure, respiratory rate, SpO₂, body temperature, sodium, potassium, creatinine, BUN, hemoglobin, and white blood cell count. Clinical notes written by physicians and nurses were also retrieved to represent the unstructured portion of the data. Together, these elements provide a detailed snapshot of each hospital stay [16].

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 [16]. Before analysis, several preprocessing steps were carried out. Missing numerical values were replaced using the mean of each feature, while missing categorical fields were filled using the most 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

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

1. **Handling Missing Values:** Missing lab values and demographic attributes were imputed using median or mean imputation, depending on variable type.
2. **Normalization:** Continuous features such as blood pressure, glucose, and creatinine levels were normalized using Min-Max scaling to ensure uniform feature distribution.
3. **Encoding Categorical Features:** Non-numeric fields such as gender or admission type were encoded using one-hot or label encoding.
4. **Feature Selection:** Redundant or low-variance features were removed to reduce dimensionality and prevent overfitting (text).

4.3.2 Training models

Once preprocessed, the structured data were transformed into numerical tensors suitable for input into deep learning models. Several machine learning and deep models were used to model structured data:

- **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 in lab time-series data.

Each model produced predictive features or class probabilities representing the structured data's 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:** Text was split into tokens using a domain-specific tokenizer.
2. **Stopword Removal:** Non-informative words (e.g., “the,” “and,” “was”) were removed using NLTK’s English stopwords list.
3. **Lemmatization:** Words were reduced to their base form using the WordNetLemmatizer to reduce lexical variability.
4. **Lowercasing and Cleaning:** All text was converted to lowercase and 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 BERT-based model pre-trained on biomedical literature (PubMed and PMC).
- **ClinicalBERT [24]:** Fine-tuned on 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 unstructured information.

4.4.3 Fusion model

The proposed model integrates two modalities:

- **Structured Data measurements:** Includes patient demographics, vital signs, and laboratory test results such as AGE, GENDER, Heart Rate, Systolic BP, Diastolic BP, Respiratory Rate, Hemoglobin, Temperature, SpO₂, Sodium, WBC, Creatinine, BUN, and Potassium.
- **Unstructured Data (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 feature-level fusion approach is widely used in multimodal deep learning because it effectively integrates heterogeneous information sources into a single representation space for improved prediction performance. The structured data and transformer text embeddings are concatenated to form a unified feature vector. This integrated representation is passed to a prediction layer that produces the final ICU mortality outcome. The concatenation-based fusion can be mathematically represented as: $Z_{\text{fusion}} = [Z_{\text{structured}} || Z_{\text{text}}]$ where $Z_{\text{structured}}$ denotes the embedding derived from tabular or numerical features, and Z_{text} represents the semantic embedding extracted from clinical notes [18]. This feature-level fusion approach is widely used in multimodal deep learning because it effectively integrates heterogeneous information sources into a single representation space for improved prediction performance. In this study, the fused representation Z_{fusion} was subsequently fed into fully connected layers for mortality outcome prediction [18].

4.5 Evaluation model

The performance of each model was evaluated using the following metrics:

1. Accuracy:

Accuracy measures the overall correctness of the model and is defined as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) [19]$$

2. Precision:

Precision represents the proportion of correctly predicted positive cases among all predicted positives:

$$\text{Precision} = TP / (TP + FP) [20]$$

3. Recall (Sensitivity):

Recall measures the ability of the model to correctly identify all actual positive cases:

$$\text{Recall} = TP / (TP + FN) [21]$$

4. F1-Score:

The F1-Score is the harmonic mean of Precision and Recall and provides a balanced measure when both metrics are important:

$$F1 = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) [22]$$

These metrics collectively measure the model's ability to correctly identify mortality outcomes and minimize false alarms.

The final layer used a Softmax or Sigmoid activation (for binary classification) to output the predicted mortality probability.

Model performance was evaluated using several standard metrics:

- **ACC:** Overall correctness of predictions
- **PRE:** Correctly predicted positive cases
- **REC:** Sensitivity to true positive cases
- **F1:** Balance between precision and recall
- **AUC:** Area under the ROC curve for robustness

Cross-validation and hyperparameter optimization (via GridSearch for traditional models and fine-tuning for deep models) were used to ensure reliability and generalization.

4.5.1 Explanation of the ROC Curve

Figure 5.2 presents the Receiver Operating Characteristic (ROC) curve for our ICU prediction model. The ROC curve shows how well the model can separate positive and negative cases at different classification thresholds [26].

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

For our model, the ROC curve remains consistently above the diagonal, indicating strong separability between classes. The Area Under the Curve (AUC) is 0.94, which demonstrates:

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

These results confirm that the model is highly effective at correctly identifying ICU outcomes while minimizing misclassifications.

4.5.2 Confusion Matrix

It is a table that summarizes how a classification model performs by comparing its predicted labels with the actual labels in the dataset. It displays four basic outcomes true positives, true negatives, false positives, and false negatives which together give a clear picture of the types of errors the model makes and how reliably it distinguishes between classes. Because it highlights both correct decisions and misclassifications, it is one of the most commonly used tools for evaluating predictive models [26].

Chapter 5 Experimental Results and Discussion

This chapter presents the outcomes of the experimental evaluation conducted using the proposed multimodal framework for ICU mortality prediction. The analysis was performed on a clinical dataset comprising two complementary data modalities: structured data, and unstructured textual data. To assess the independent contribution of each data modality, a range of machine learning and deep learning models were trained separately using the structured 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 standard classification metrics, including Accuracy, Precision, Recall, F1-score, and AUC.

5.1 Experimental Setup

The experiments 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.

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 F1-score among the models.

Models	Accuracy	Precision	Recall	F1-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 accuracy, precision, recall and F1 score. Applying TF-IDF improved the way the model emphasized important clinical terms. Logistic Regression and SVM both did well again, each scoring 89.94% accuracy, while tree-based models lagged behind. This shows that linear models handle high-dimensional, sparse data better. Decision Tree had the lowest accuracy at 79.25%, probably because it overfitted and couldn't generalize across the sparse feature space. Using contextual embeddings had a notable impact on model performance. BioBERT achieved an accuracy of 84.08% with an F1-score of 80.62%, while ClinicalBERT performed slightly better, reaching 85.35% accuracy and an F1-score of 81.30%. ClinicalBERT also showed the highest precision at 98.04%, highlighting its ability to correctly identify positive mortality cases. However, its recall was 69.44%, suggesting that some positive cases were missed despite the high reliability of its predictions. This balance between precision and recall reflects a conservative classification strategy, which can be advantageous in clinical risk prediction, where minimizing false positives is important. The better performance of **ClinicalBERT compared to BioBERT** is likely due to its domain-specific pretraining on clinical notes, giving it a deeper understanding of medical terminology and context. Among the classical models, LR performed best, particularly with Word Count and TF-IDF features. For deep contextual models, ClinicalBERT achieved the highest precision and a competitive F1-score, showing its strength in capturing subtle linguistic patterns in clinical text. Overall, these results indicate that while traditional text representations with simple models remain solid baselines, domain-specific transformer models like ClinicalBERT offer improved precision and interpretability for healthcare prediction tasks.

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

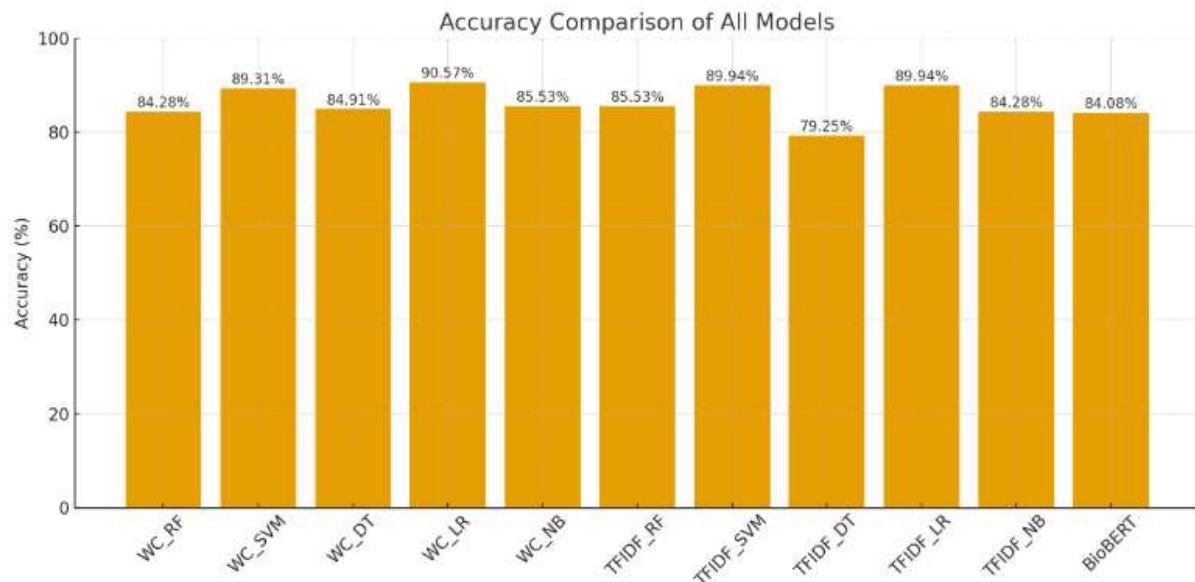


Fig 5. 1 Accuracy comparison of all models

5.5 Results of fusion model

The Fusion model achieved the best overall results among all the approaches tested, reaching 91.08% accuracy, 91.13% precision, 91.08% recall, and an F1-score of 91.09%. These results highlight the benefit of combining structured data, such as laboratory tests and vital signs, with unstructured clinical notes in a single deep learning framework.

Unlike models that rely only on text or numerical data, the multimodal fusion approach takes advantage of complementary information from both sources. Its Transformer-based design allows

the model to capture complex relationships between physiological measurements and the language used in clinical documentation. This produces richer patient representations and helps the model better distinguish between high-risk and low-risk ICU cases.

The similar values across precision, recall, and F1-score indicate that the model is both sensitive correctly identifying mortality cases and specific avoiding false alarms. This balance demonstrates that the Multimodal Transformer generalizes well and can robustly detect clinically relevant patterns.

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, 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.

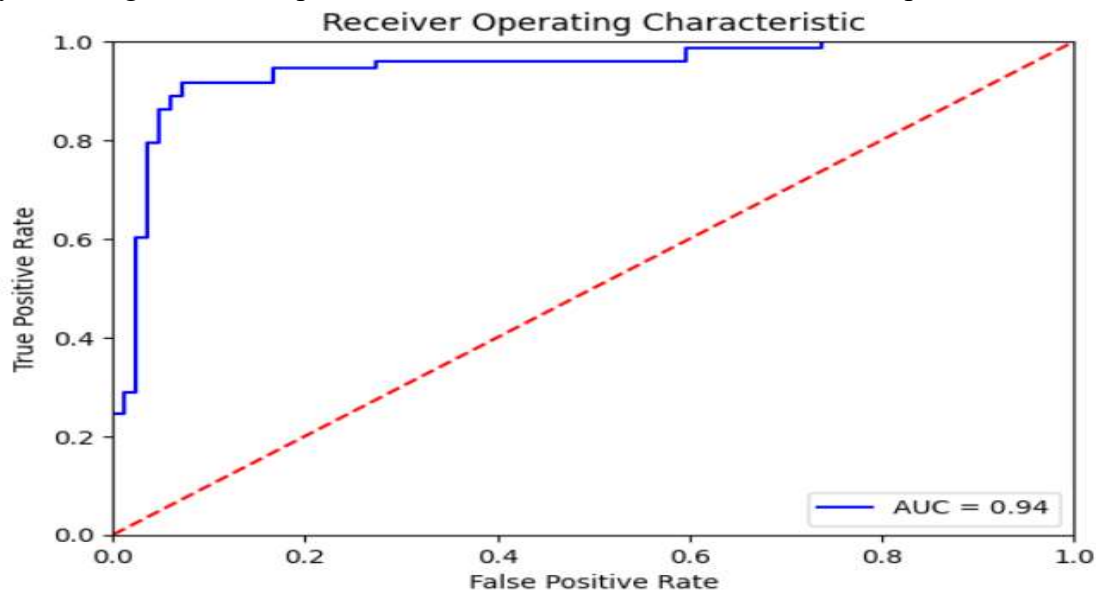


Fig 5. 2 Receiver Operating Characteristic (ROC) curve of the model

In the figure 5.3, The confusion matrix 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.

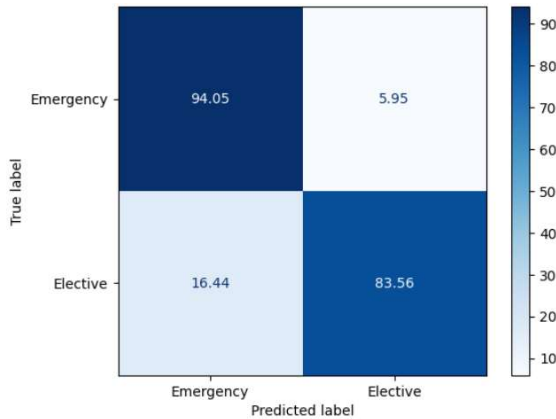


Fig 5. 3 Confusion Matrix

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

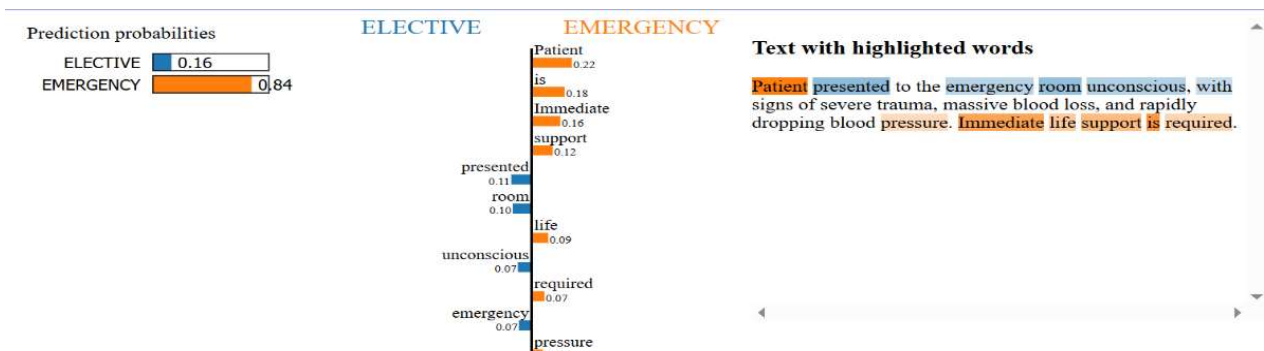


Fig 5. 4 Xai Emergency Case

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

- **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

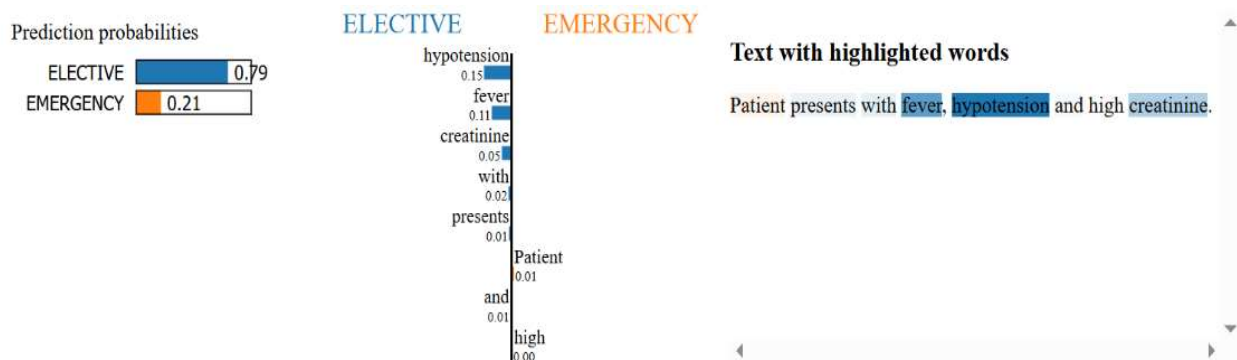


Fig 5. 5 Xai Elective Case

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 structured data, or other parts of the note that carry weaker influence, outweighed those warning terms.

5.7 Discussion

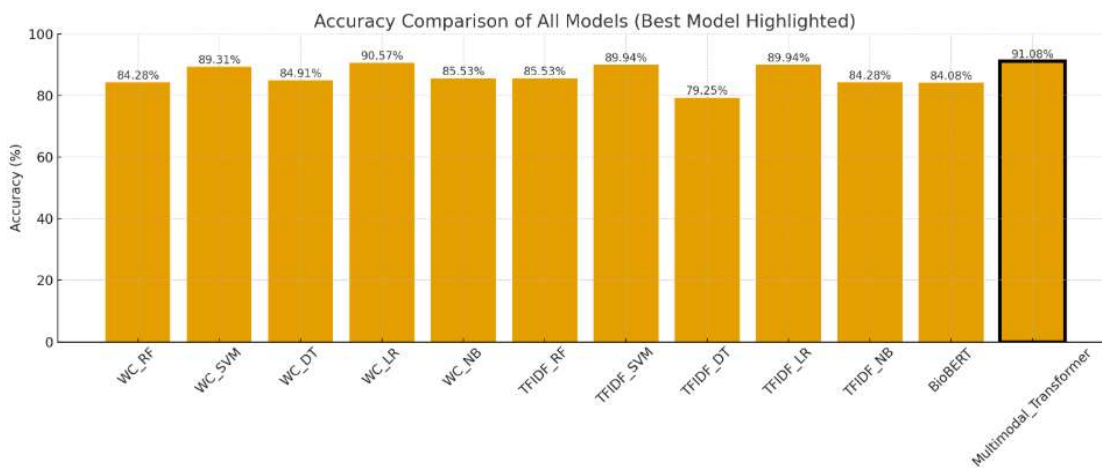


Fig 5. 6 Accuracy comparison of all models, highlighting the best model

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. By integrating its embeddings with structured clinical features through the Transformer fusion network, the model captures both semantic and physiological aspects of patient health. As a result, the Multimodal Transformer Fusion model offers a more complete view of ICU patient status, delivering the most accurate and balanced mortality predictions.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This research set out to improve ICU mortality prediction by combining structured clinical data with unstructured text from patient notes. Instead of relying on a single type of information, the proposed multimodal framework brings together two complementary sources: numerical measurements such as laboratory results and vitals, and contextual information extracted from clinical narratives. BioBERT and ClinicalBERT were used to capture the meaning and subtle patterns within the text, while traditional machine-learning and deep models handled the structured variables. The fusion of these two streams created a more complete view of each patient's condition. Across all evaluation metrics precision 91.13, accuracy 91.08, recall 91.08, F1-score 91.09, and AUC 95 the multimodal model consistently performed better than models trained on structured data alone or text alone. These results highlight the value of using domain-specific transformer models to understand clinical language and pairing them with established algorithms for numerical features. Overall, the study demonstrates that multimodal learning is not only feasible in critical-care settings but also adds meaningful improvements in predictive accuracy. This integrated approach can support clinicians by offering earlier and more reliable insights into patient risk, which may contribute to faster interventions and better outcomes in the ICU.

6.2 Future Work

Future work can focus on

- 1- expanding the dataset, incorporating additional clinical modalities such as medical imaging, and exploring more advanced fusion strategies that go beyond simple concatenation.
- 2- Improving model interpretability is another important direction, allowing clinicians to better understand the factors influencing predictions.
- 3- Testing the framework across multiple hospitals and real-time ICU settings would help validate its robustness and support its translation into practical clinical decision-support tools.

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.