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

مجلة ميدأوشن للبحوث والدراسات

ISSN: 3078-5669

Hizam Ahmed Al saeedi

Innova Medical Services and Applications Company
Artificial Intelligence Researcher

Received: 29/10/2025
Reviewed: 23/12/2025
Issued at: 15/1/2026

Abstract

Objectives: This study presents a multimodal artificial intelligence framework aimed at improving mortality prediction in intensive care units (ICUs) by integrating structured clinical data with unstructured clinical notes. The framework combines traditional machine learning algorithms with transformer-based language models to capture both numerical patterns and nuanced textual information recorded during patient care. For structured data analysis, machine learning models including Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Naive Bayes (NB) were employed using laboratory measurements, vital signs, and demographic features.

Methodology: To process unstructured textual data, the study fine-tuned BioBERT and ClinicalBERT models, which are specifically designed to interpret medical language and are pre-trained on large-scale clinical corpora. These models transform clinical narratives into meaningful contextual representations. In parallel, structured variables, including laboratory results and vital signs, were processed independently to generate complementary feature representations.

Results: The findings demonstrate that integrating clinical text with structured data significantly enhances predictive performance compared to using either data source independently. The combination of transformer-based language models with machine learning techniques and self-attention mechanisms contributes to more robust and reliable mortality prediction.

Conclusion: The proposed framework provides a practical and scalable foundation for clinical decision support systems and shows strong potential for improving risk assessment and patient management in intensive care settings.

Keywords: Multimodal, ICU, Mortality, Xai, framework.

Introduction

Predicting patient status in the ICU is one of the most challenges that facing many ICU management in clinical decision-making. ICU patients often has complex medical conditions that progress rapidly, requiring from clinicians to make quick decisions (Ruan et al., 2025) Traditional methods use a single type of data, often mainly structured data such as research facility results and vital signs. The structured data provides valuable data but not enough, it does not cover all the

important information found in unstructured clinical records, such as physician notes, progress reports, and discharge summaries (Gao et al., 2024)

A lot of ICUs departments are not used unstructured clinical records with structured data 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 (Ruan et al., 2025).

This paper focused on how to build multimodal AI framework to improve the accuracy of ICU fatality estimation by marge structured data and unstructured data, by use Machine Learning (ML) methods, natural language processing (NLP) and deep learning (DL) to enhance predictive performance and facilitate the interpretation of clinical outcomes.

Research Problem

Most current models for predicting patient mortality in ICU rely primarily on structured clinical data, neglecting valuable information from clinical observations. This deficiency leads to lower prediction accuracy and overlooks early indicators of patient deterioration, highlighting the need for an integrated, multimodal data approach.

Study Objectives

Primary Objective: Develop a multimodal fusion framework that integrates structured clinical data (vitals, labs, demographics) with unstructured clinical notes using domain-specific transformers (BioBERT/ClinicalBERT) for ICU mortality prediction.

Performance Enhancement: Achieve superior prediction accuracy compared to single-modality approaches, targeting >90% AUC and balanced precision-recall metrics for reliable clinical decision support.

Interpretability & Validation:

Implement explainable AI techniques to provide transparent decision-making insights and validate the framework against using established ICU datasets.

Previous Studies:

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 structured data only, while some studies use clinical notes only, and some try to combine both. (Gao et al., 2024)

ICUs department prediction mostly used structured data, 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 (Saleh et al., 2024)

Gao et al., (2024) 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) 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) 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) 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) 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., (2021) 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.

Ruan et al., (2025) was built a system that can marge two type of data, structured data 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) The author used CNNs on nursing notes. The ROC-AUC was 0.87. 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. This step is especially critical for healthcare data, where there are many features, complex interactions, and occasional inconsistencies or noise.

When comparing 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 (Gao et al., 2024) applied LGBM, DT, RF, SVM, XGB and MLP and RF was the best model with 85.99. In (Ruan et al., 2025) the authors applied MLP, ResNet and FT-Transformer and the FT-Trans achieved the high score with 87.32. In (Huddar et al., 2016), the authors applied LR, SVM, DT, AdaBoost and RF and the best model was SVM with 88.58 score. In (Xu et al., 2019), the authors applied Text-TF-IDF-CNN, LS, DR and TD to achieved 76.33 score. The authors applied (Lyu et al., 2023) 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 (Shukla et al., 2020) applied CNN, TF-IDF and RNN and the TF-IDF / 1-NN was the best with score 86.27. In (King et al., 2023), 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.

Table (1): Comparison with Literature Studies

Models	Dataset	The best model	Result
LGBM, DT, RF, SVM, XGB, MLP	MIMIC-IV	RF (Cross Validation)	85.99
MLP, ResNet, FT-Transformer	MIMIC-IV	FT-Trans	78.32
LR, SVM, DT, AdaBoost and RF	MIMIC II	SVM	88.58
CNN, LS, DR, TD	MIMIC-III	CNN	76.33
Transformer fusion, Clinical BERT, BioBERT , BERT	MIMIC-III + eICU-CRD	Clinical BERT	85.10
CNN, TF-IDF, RNN	MIMIC-III	TF-IDF / 1-NN	86.27
Self-supervised contrastive + LSTM, LR, Baseline	MIMIC-III	Baseline	85.5
RNN, SVM, DT, RF, NB, LSTM, Transformer, Clinical BERT, BioBERT	MIMIC-III	Fusion	91.13

5. Methodology

Introduction

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 (Ruan et al., 2025)

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

It consists of six key stages:

- Data collection
- Data Preprocessing for SD and UD
- Feature representation
- Modal Training
- Fusion modal
- Classification result

5.2 Data

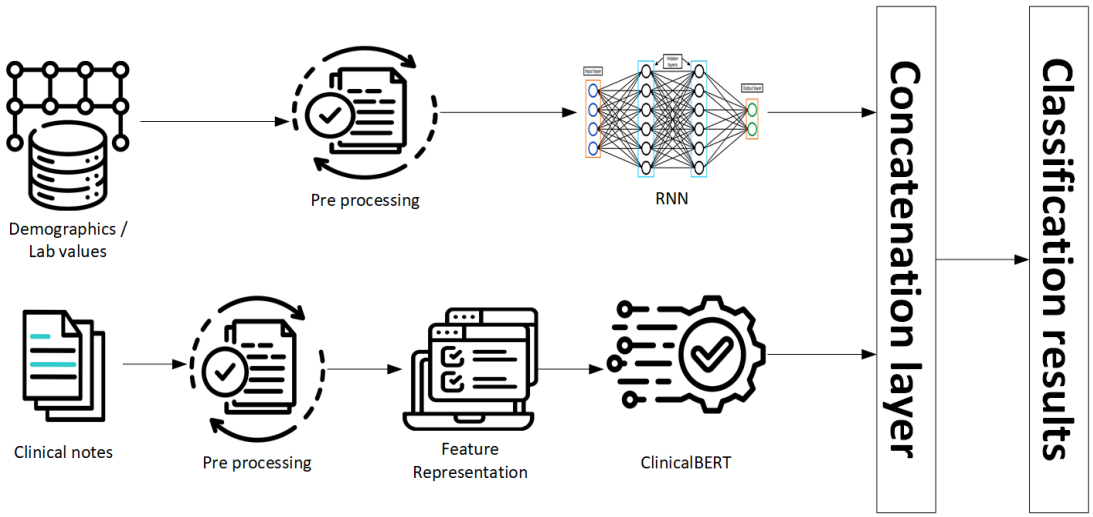

Fig. (1): 1 ICU system architecture

Collection

The dataset used in this paper was download from the MIMIC-III (Johnson et al., 2026) . This dataset contains of two types of data, structure and unstructured data, 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 types of data. The part one is structured data includes demographic details like age, gender, vital sign and laboratory result. For example, some features like heart rate, systolic and diastolic blood pressure, respiratory rate, SpO_2, body temperature 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 (Saleh et al., 2025). 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 (Lee et al., 2020).

Structured data

Data Preprocessing

Structured data are often noisy, incomplete, and heterogeneous in ICU settings. To ensure data quality and consistency, the subsequent preprocessing steps were implemented (Saleh et al., 2025):

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

Training models

Once preprocessed, the structured data were transformed into quantitative data matrices proper for input into deep learning 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 structured data's contribution to the overall mortality risk.

Unstructured data

Text Preprocessing

Clinical notes contain rich contextual information but require careful preprocessing to handle linguistic noise (Saleh et al., 2025).

The following steps were performed:

- **Tokenization:** We are split text into tokens using a domain tokenizer.
- **Stopword Removal:** Remove words that not have value in context (e.g., "the," "and," "was") from stopword list.
- **Lemmatization:** Return word to the base.
- **Lowercasing and Cleaning:** Converted test to lowercase and remove punctuation, digits, and special characters were removed.

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 frequency a specific word emerges 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 enhances 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 highlight clinically meaningful terms while reducing the impact of common but uninformative words (e.g., "patient," "the") (Saleh et al., 2025).
- **BioBERT (Lee et al., 2020) :** A This model was pre-trained on biomedical literature (PubMed and PMC).
- **ClinicalBERT** This model has training from hospital discharge summaries and MIMIC-III (Johnson et al., 2016) 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 ultimate concealed states from the [CLS] symbol was extracted as the textual feature vectors representing the patient's unstructured information.

Fusion model

The fusion model integrates two models:

- **Structured Data:** Patient demographics, vital signs, and laboratory result.
- **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 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 + unstructured data.

Evaluation model

We are evaluating our model with this metrics:

- Accuracy: Accuracy measures how well a model is generally correct, and is defined as: (TP + TN) / (TP + TN + FP + FN) (Liu et al., 2022)
- Precision: Precision represents the number of positive cases out of all positive cases : TP / (TP + FP) (Merchant et al., 2024)
- Recall (Sensitivity): Recall represents the number of positive cases divided by total of positive cases and false negative: TP / (TP + FN) (Dritsas & Trigka, 2025)
- F1-Score: The F1-Score is depended of P and R to get the best evaluation for our model :
- F1 = (2 × P × R) / (P + R) (Hamze & Klarmann, 2024)

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:** Area under the ROC curve for robustness
- **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

Explanation of the ROC Curve

Figure 4.2 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 (Fawcett, 2006) .

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

Confusion Matrix

The CM table demonstrates the performance of a classification 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.

Results and Discussion:

Introduction

This chapter presents our result and experimental setup for our model. 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.

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.

Results of structured data

Table (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.

Table (2): Comparison structure 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

Results of unstructured data

Table (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. The better performance of ClinicalBERT was achieved good result comparing with BioBERT and The Precision was the highest score for all other models.

Table (3): Comparison of All Models

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

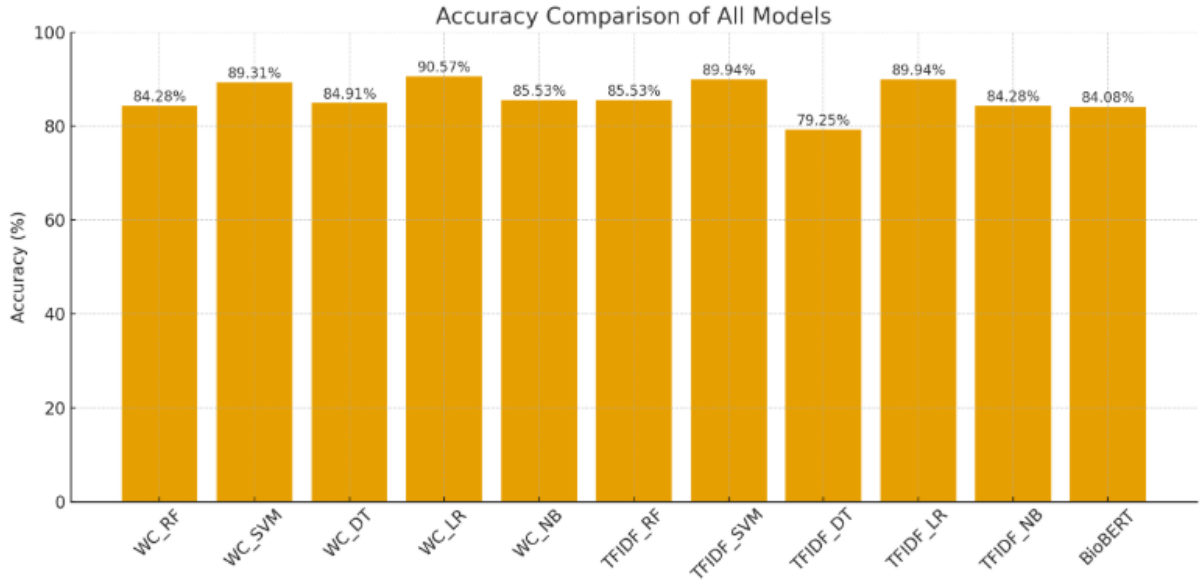

Fig. (2): Accuracy comparison of all models

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 fine tuning model and other models by transformer design.

Table (4): Comparison of Multimodal

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

At the next figure, 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 selected beneficial instance greater than a negative one 94% of the time, which is considered very good performance in clinical prediction tasks.

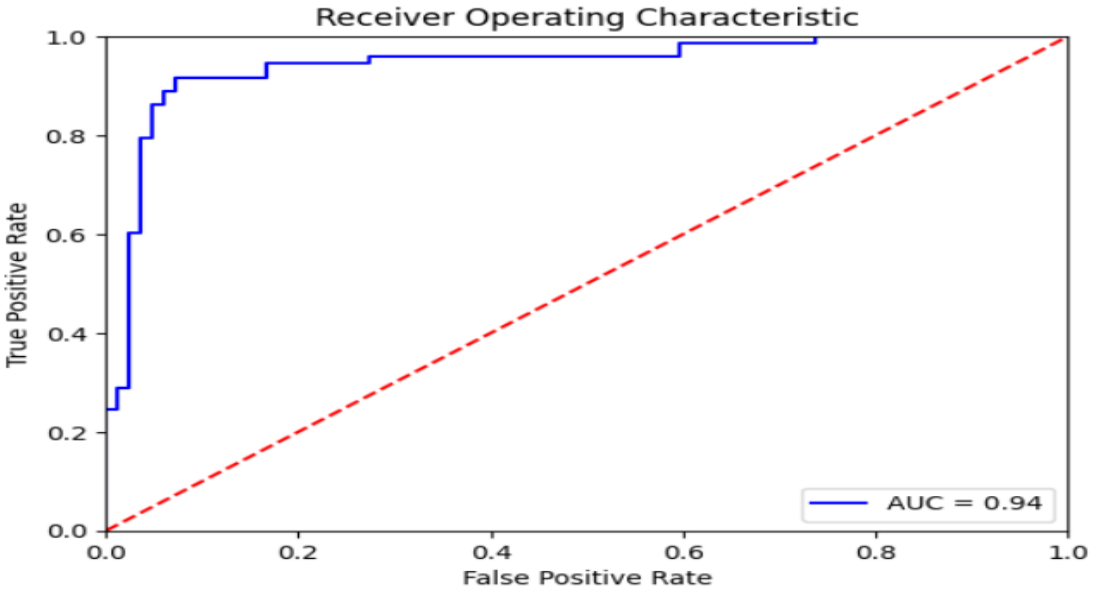

Fig. (3): Receiver Operating Characteristic (ROC) curve of the model

In the figure (4), 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.

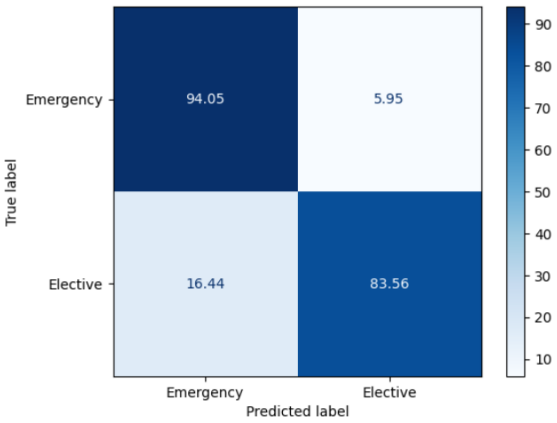

Fig. (4): Confusion Matrix

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 6.4 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

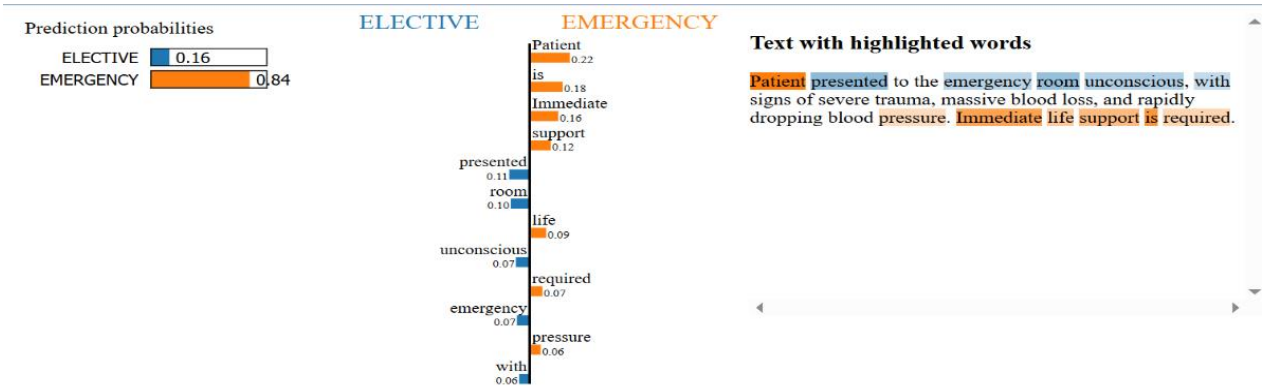

Fig. (5): Xai Emergency Case

In the notes even when some of those words are very common because their meaning

ELECTIVE case, Figure (6) shows a word-by-word explanation for one specific prediction. Even though the clinical note includes several expressions that normally point to an acute 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.

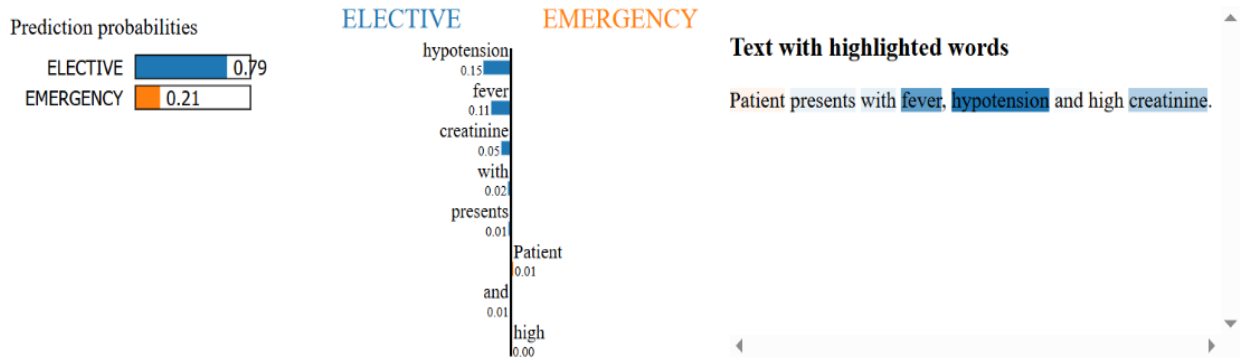

Fig. (6): Xai Elective Case

Discussion

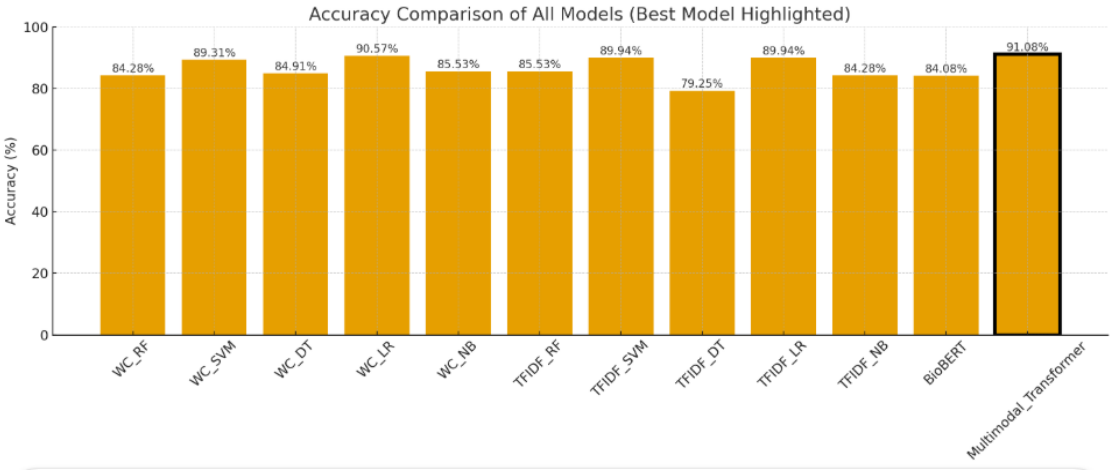

Fig. (7): Accuracy comparison of all models, highlighting the best model

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.

Conclusion

This paper improves the accuracy of ICU mortality prediction by integrating organized information and unstructured data 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 outcomes and essential signs, and medical annotations. The BioBERT and ClinicalBERT models were used to extract meaning from clinical notes, in addition DL models handled structured. 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 paper demonstrates that multimedia learning has proven its predictive power in critical care settings and significantly improves prediction accuracy.

References:

Gao, J., Lu, Y., Ashrafi, N., Domingo, I., Alaei, K., & Pishgar, M. (2024). Prediction of sepsis mortality in ICU patients using machine learning methods. *BMC Medical Informatics and Decision Making*, *24*(1), 228.

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. *Journal of Healthcare Informatics Research*, 1-42.

Huddar, V., Desiraju, B. K., Rajan, V., Bhattacharya, S., Roy, S., & Reddy, C. K. (2016). Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, *4*, 7988-8001..

Al-Dailami, A., Kuang, H., & Wang, J. (2025). Multimodal Representation Learning Based on Personalized Graph-Based Fusion for Mortality Prediction Using Electronic Medical Records. *Big Data Mining and Analytics*, *8*(4), 933-950. Iwase, S., Nakada, T. A., Shimada, T., Oami, T., Shimazui, T., Takahashi, N., ... & Kawakami, E. (2022). Prediction algorithm for ICU mortality and length of stay using machine learning. *Scientific reports*, *12*(1), 12912.

Yang, H., Kuang, L., & Xia, F. (2021). Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, *12*(1), 3.

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.

Kumar, S., Oh, I., Schindler, S., Lai, A. M., Payne, P. R., & Gupta, A. (2021). Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. *JAMIA open*, *4*(3), ooab052.

Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Xu, K., Lam, M., Pang, J., Gao, X., Band, C., Mathur, P., ... & Xing, E. P. (2019, October). Multimodal machine learning for automated ICD coding. In *Machine learning for healthcare conference* (pp. 197-215). PMLR.

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.

Lyu, W., Dong, X., Wong, R., Zheng, S., Abell-Hart, K., Wang, F., & Chen, C. (2023). A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings* (Vol. 2022, p. 719).

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*.

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.

Saleh, H., McCann, M., El-Sappagh, S., & Breslin, J. G. (2024). Transformer Fusion Net: A Real-Time Multimodal Framework for ICU Heart Failure Mortality Prediction Using Big Data Streaming. In *2024 International Conference on Computer and Applications (ICCA)* (pp. 1-6). IEEE.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*(1), 1-9.

Liu, S., Wang, X., Hou, Y., Li, G., Wang, H., Xu, H., ... & Tang, B. (2022). Multimodal data matters: Language model pre-training over structured and unstructured electronic health records. *IEEE Journal of Biomedical and Health Informatics*, *27*(1), 504-514.

Merchant, A. M., Shenoy, N., Lanka, S., & Kamath, S. (2024). Ensemble neural models for ICD code prediction using unstructured and structured healthcare data. *Heliyon*, *10*(17).

Dritsas, E., & Trigka, M. (2025). Applying Machine Learning on Big Data with Apache Spark. *IEEE Access*.

Hamze, H., & Klarmann, S. (2024). Implementing Apache Kafka in industrial environment to enable data streaming for cloud-based applications. In *Dalam Proceedings of the 7th European Conference on Industrial Engineering and Operations Management (hlm. 278–286). IEOM Society International. https://doi. org/10.46254/EU07* (Vol. 20240082).

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*(1), 1-9.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.