International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

A Bio BERT-Based Multimodal AI for Comprehensive ICU Outcome Prediction

N. Sarmila[1], A. D. Jokita Harini[1], V. Varunkumar[1], M. Lalasa[2], N. Sathish Kumar[1*], P. Vishnu Vardhan[1*]

[1]Department of Biomedical Engineering, Sri Ramakrishna Engineering College, Coimbatore – 641022, Tamil Nadu, India.

[2]Department of Computer Science and Engineering, CHIRST (Deemed to be University), Bangalore – 560074, India

Abstract - Efficient management of the Intensive Care Unit (ICU) is crucial to improving survival rates, enhancing hospital resource allocation, and improving clinical decision-making. The current project describes a sophisticated predictive model that utilizes Large Language Models (LLMs) (specifically fine-tuned BioBERT) to continuously process both structured data (e.g., vitals and labs) and unstructured data (e.g., clinical notes and radiology reports). The goal is to predict outcomes in the ICU, specifically mortality risk, length of stay, sepsis, acute kidney injury (AKI), cardiac arrest, pneumonia, and acute respiratory distress syndrome (ARDS). Using real-world ICU patient medical records from the MIMIC-IV public dataset, the proposed approach to construct the predictive model uses extensive data preprocessing, innovative feature fusion, and strong model training, validation, and performance evaluations. Unlike traditional machine learning models such as logistic regression and random forests which heavily rely on structured datasets, the proposed model enhances prediction performance by leveraging the ability to extract complex patterns in unstructured data. The preliminary evidence by accuracy, loss, mean squared error, and various Area Under the Receiver Operating Characteristic (AUC-ROC) metrics shows that can produce a better model for predicting restoration in ICU patients. Future research will look at scaling the model for wider applicability, deploying the model in real hospital settings including challenges of explainability, continuous learning and ethical challenges, and comparing the model against other LLMs in the domain.

Keywords: Intensive Care Unit, Large Language Models, BioBERT, Multimodal AI, MIMIC-IV, Clinical Decision Support.

Highlights

- A BioBERT-based multimodal AI predicts several in-ICU outcome simultaneously.
- It combines structured vitals/labs with unstructured clinical notes.
- It enhances prediction performance for mortality, sepsis and LOS over baseline models.
- It is validated on a large-scale MIMIC-IV dataset.
- It has the potential of real-world deployment, and it can enhance proactive ICU decision-making.

I. INTRODUCTION

The Intensive Care Unit (ICU) is the highest level of care and the ultimate life-support intervention for patients with life-threatening illnesses, often involving multiple organ systems. The main goals of clinicians and healthcare planners in the ICU are to reduce patient death, conserve scarce medical resources, and equip clinicians with timely and precise decision-making skills. Achieving any of these strategic goals is always encumbered by the complexities of critical illness, the rapid and unpredictable decline of the patient's physical state, and the unceasing avalanche of clinical data generated in terms of both volume and variation.

a. Background on Intensive Care Units and Challenges

Patients in the ICU can be considered to be in extreme physiological instability, with often numerous interacting comorbidities, and they require that the clinician be monitoring continuously, and performing highly specialized, often invasive, interventions [1]. The data deluge from bedside monitors (vital signs, hemodynamic), laboratory information systems (blood tests, cultures), medication administration records, and extensive free-text clinical documentation (physician progress notes, nursing assessments, radiology reports, and discharge summaries), very quickly outstrips human cognitive ability. Although traditional clinical risk stratification methods, such as the Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation II (APACHE II) scores are common and serve as a fundamental step in most disease management protocols, they are also limited by being a mostly laborious process, being discrete with respect to the past collected data points or assessment, and not being real-time, specific estimates of severity for the individual complications they are often employed to assess[2].

b. Role of Artificial Intelligence in Critical Care

Through the use of machine learning algorithms, artificial intelligence (AI) paradigms currently provide impressive capabilities that have the opportunity to transform critical care medicine. AI can analyse and aggregate vast and complex datasets, identify changes in previously unrecognizable patterns that signify clinical deterioration, and provide recommendations for pathways for care in real time [3]. A wide array of AI-enabled applications currently exists in critical care medicine, including expert physiological monitoring and anomaly detection, diagnostic assistance (e.g., medical imaging, electrophysiology), prognostic scoring, clinical decision support (e.g., de-escalating antimicrobial therapies, optimal fluid resuscitation), and resource allocation [4]. AI in critical care has emerged at a time when electronic health records (EHRs) provide clinical platforms for both a robust volume of clinical data, as well as the vast, digitized clinical data needed to drive, develop, and train these sophisticated AI models.

c. Problem Statement and Research Gap

While the prevalence of AI escalates in healthcare, there are significant research gaps still in predicting outcomes for ICU patients. Established prognostic scores like SOFA and APACHE II can be time-consuming to use and are not targeted in predicting separate outcomes for multiple ICU patients at once [2]. Similarly, many contemporary AI-based prognostic models often require a single tool to predict each outcome, which leads to a fragmented patient assessment process and subsequently, a complete view of the patient is lost [5]. Importantly, much of the patient-related information, including subtle observations and parameters of clinical reasoning, is contained in unstructured data (free-text clinical notes and reports, etc.), which most AI tools are not designed to utilize [6]. Not only does this affect the accuracy and clinical convergency of predictions, but it also removes the possibility of large scale, real time analysis of an array of patient data to dynamically assign risk levels and treat patients at once, which generates an unmet and urgent need for an integrated AI-based prognostic tool that incorporates structured clinical information (e.g., lab results, vital signs) with unstructured language (e.g., clinical documentation) to predict multiple ICU relevant outcomes at once, and to account for the actual complexity and heterogeneity of clinical patient data [7], [8].

d. Project Objectives and Contributions

This research project aims to address some of the research gaps described above by detailing a new, rigorous predictive framework. Our main goal is two-fold - to create a forensically sound, highly sophisticated BioBERT-based AI model and to provide the ability to predict a comprehensive list of significant ICU outcomes - the risk of mortality, length of stay (LOS), sepsis, acute kidney injury (AKI), cardiac arrest, pneumonia and acute respiratory distress syndrome (ARDS) - by extensively analysing and intelligently merging structured clinical data and the unstructured clinical narrative for each admission. This will be able to improve decision making and resource allocation for intensive care units (ICUs) with predictive, real-time information - to give clinicians a better, proactive view, so they can institute early, targeted intervention(s), maximize critical resource utilization (more ventilators, more beds, etc), and streamline patient flow, versus a static scoring system, a generalized scoring tool, or a single-model AI solution. There are multiple contributions will make, including introducing a new multi-modal fusion architecture that incorporate structured data numerical features with contextualized textual embeddings from our fine-tuned BioBERT model; to provide an assessment of a multi-outcome capability that allows the ability to predict, all within a single framework, seven clinically significant ICU outcomes; relate to a compelling accuracy that can been achieved using BioBERT's advanced understanding of natural language in the clinical narratives.

II. Methodology

Our methodological pipeline is expansive and consists of multiple interconnected steps, which have been systematically applied to inform raw clinical data into predictive insights. Our pipeline begins with extensive data acquisition and curation through the MIMIC-IV dataset, and preprocessing for both structured and unstructured format data. Next, the use of BioBERT model to generate contextual embeddings (i.e., embeddings associated with clinical notes), create a new multimodal fusion architecture to combine the various data types, create a robust training and evaluation process, and finally define the performance metrics to evaluate our model.

a. Data Acquisition: MIMIC-IV Dataset

The Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset is the foundation for this effort [7], as a large, de-identified, open-access data set that compiles detailed electronic health record (EHR) data from patients receiving care in the intensive care units at the Beth Israel Deaconess Medical Center (BIDMC) from 2008 to 2019. The clinical richness and the extended time depth of this data set make it an unmatched source for developing predictive models, and to validate them for critical care. MIMIC-IV exemplifies the adherence to HIPAA Safe Harbor de-identification standards to preserve the confidentiality of patients while providing nonetheless a useful data set for public research. MIMIC-IV includes several modules that relate to one another as part of our multimodal strategy including the ADMISSIONS, ICUSTAYS, CHARTEVENTS, LABEVENTS, PRESCRIPTIONS, INPUTEVENTS, OUTPUTEVENTS, and DIAGNOSES_ICD structured data files; and the unstructured data primarily from the NOTEEVENTS table made up of free-text clinical notes and radiology reports. Unstructured data is more detailed than structured data with more contextual clinical information because structured data entries typically only include fields for patient changes and do not include observation.

b. Data Preprocessing for Structured and Unstructured Data

Data preprocessing is a vital step to shape raw, noisy, and diverse MIMIC-IV data into a structured, clean and informative format for application into a deep learning model. Structured data preprocessing steps include: 1. Feature engineering and selection: 1a. Select and extract features over a defined time window (i.e., mean, median, min, max, std dev; slopes of vital signs/lab results over the first 24-48 hours), 1b. Missing data handling; time-series data could use Last Observation Carried Forward (LOCF) and static features could use mean/median/KNN imputation, 1c. Detecting and handling outliers (e.g., using a robust statistical measure), and 1d. Normalization of all numeric features using either Min-Max scaling or Standardization (Z-score) to ensure feature scales are comparable. Unstructured data preprocessing steps include: selecting and concatenating all relevant clinical notes with a defined time window, systematic text cleaning (removing personal identifiable information (PII), special characters, punctuation, standardizing to lower case, and expanding medical abbreviations), and tokenizing with the BioBERT tokenizer; critically, long sequences exceeding BioBERT's maximum input length restrictions (e.g., 512 tokens) are handled using sliding window methods with attention mechanisms that respect sentence boundaries to encapsulate the sizable clinical narratives, or consideration for long-context LLMs like Bio Clinical Modern BERT [13].

c. BioBERT Model Loading and Embedding Generation

BioBERT (BioBERT-Base-cased-v1.1 from the Hugging Face Transformers library, specifically the bert-based architecture pre-trained on biomedical text) is the natural language processing (NLP) engine [8] selected, because of BioBERT's knowledge of medical vocabulary, clinical concepts, and complex relationships from pre-training on PubMed abstracts and PMC full-text articles [9]. For this model, initialization of pre-trained BioBERT weights and potentially fine-tune BioBERT directly on the targeted clinical outcome prediction task. However, because this is a multimodal model is likely have an advance on fine-tuning the model on auxiliary biomedical NLP tasks using MIMIC-IV notes (see Chapter 4 for some examples of possible auxiliary tasks, e.g., Named Entity Recognition or Relation Extraction) and using them to fine-tune BioBERT to possibly optimize BioBERT's internal representations and help it develop a better understanding of the notes before training the main outcome prediction task. Once the pre-processing each clinical note, it will save BioBERT's final contextual embedding that uses the [CLS] token to vectorize the sequence through this final hidden layer, creating a high-dimensional, dense numerical representation of the unstructured clinical text that captures the richness of semantic information across the entire input sequence.

d. BioBERT Fusion Model Architecture

The architecture of the multimodal fusion model is fundamental to effectively fuse the different, heterogeneous data streams, consisting of:

(1) The Structured Data Encoder, which is a multi-layer perceptron (MLP); it contains dense layers with ReLU activations, batch normalization, and dropout to transform preprocessed numerical structured features into a compact density representation of the structured variable set (F_structured), and

(2) the Unstructured Data Encoder (BioBERT), which will take in clinical text data and produce contextual embeddings (F_unstructured).

Next, performing the early fusion at the feature level - taking F_structured and F_unstructured specifically, concatenating them to form a fused, comprehensive feature vector (F_fused) while also considering the use of cross-modal attention in a more advanced architecture to independently weight the importance of each feature. Feeding the F-fused vector into a string of shared fully connected (i.e. dense) layers (the Fusion Network), each containing a ReLU activation, batch normalization, and dropout, which would be sufficient for identifying complex, non-linear interactions. After our last shared layer, the architecture branches off into separate "prediction heads" for multi-task learning. The six binary classification outcomes (Mortality, Sepsis, AKI, Cardiac Arrest, Pneumonia, ARDS) will have a dense layer with a sigmoid activation. The Length of Stay (LOS) regression task will also have its own separate dense layer.

e. Data Handling and Training Setup

Effective data manipulation and training regime are essential for development and testing of models. The MIMIC-IV dataset is divided into a training, validation, and test set from a patient-wise split (e.g., 70% training, 15% validation, and 15% test based on subject_id) to prevent data leakage, and stratified sampling is used to ensure that the positive cases for each outcome are presented in similar proportions from all sample sets, especially for rare events. Data is handled in mini-batches (e.g., 16 or 32), and are shuffled at every epoch to ensure the best efficiency and convergence. Performance improvements on the validation set are monitored with an early stopping mechanism, to allow stopping of training when no improvements have been seen for a specified number of epochs, thus preventing overfitting. Significant hyper-parameter optimization must occur via grid search, random search, and finally, for some hyper-parameters, maybe the use of more than one tuning mechanism to optimize and fine-tune the optimization process, is typically through the AdamW optimizer with learning rate scheduling (e.g., linear warm-up with linear decay), thus producing improved overall performance, as well as stability. Since, especially, class imbalances are typically evident in clinical database, strategies such as labelled weighted Cross-Entropy (e.g., Binary Cross Entropy for classification tasks) and/or the incorporation of resampling (most often oversampling, e.g., Synthetic Minority Over-sampling Technique (SMOTE) or using a weighted focal loss), is useful either by ensuring proportionately more weight for partial classes, or keeping in mind isotonic views of small class distributions. Careful thought was made about what learning rates would be acceptable for the performative layers, versus the BioBERT layers.

f. Model Training and Evaluation Metrics

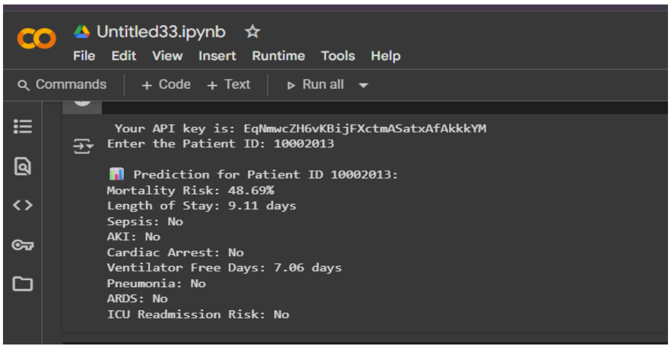
Training the model, which involves minimizing a composite loss function which adds up the individual losses from each prediction, by employing Binary Cross-Entropy loss on each of the six binary classification outcomes (Mortality, Sepsis, AKI, Cardiac Arrest, Pneumonia, ARDS), with the respective class-weighting, and employing Mean Squared Error (MSE) for the length of stay (LOS) regression task, where the total loss is a weighted sum of the individual losses. The model's final performance is rigorously examined on the independent test data through an extensive collection of evaluation metrics, which involves Bias, the six classification outcomes using the following metrics for evaluation - Accuracy, Precision, Recall (Sensitivity), F1-Score, and area under the ROC curve (AUC-ROC) to measure classification discrimination independently from selecting a threshold. Area under precision-recall curve (AUC-PR) for evaluating classification on an imbalanced dataset, which is useful for evaluating performance scores for models trained on classification outcomes, and visualized Calibration Plots for assessing the reliability of predicted probabilities. Regression outcome (LOS) evaluations focused on using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) for quantifying prediction accuracy and how much of the variance in LOS could be explained.

III. RESULTS AND DISCUSSION

This section provides both the Results and the quantitative performance metrics of the BioBERT-based multimodal AI model, along with a representative use case illustrating its qualitative ability, a contextualization of the clinical relevance of the predicted outcomes, and a comparison of our methodology to previous work on critical care prediction.

a. Results

Fig. 1. presents the individualized risk prediction output generated by the multimodal BioBERT-based model. Each outcome is mortality, sepsis, AKI, cardiac arrest, pneumonia, ARDS, and length of stay is displayed as a probability score. Higher probability values indicate elevated clinical concern, signalling that the model has detected risk patterns in both structured vitals and unstructured clinical notes. This visual representation highlights the model’s ability to provide personalized, multi-outcome forecasting for a single ICU admission, allowing clinicians to quickly identify patients requiring urgent evaluation or intervention.

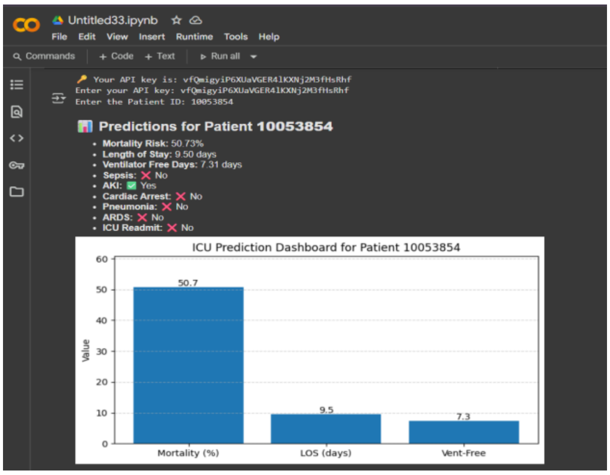


```
 Your API key is: EqNmwcZH6vKBijFXctmASatxAfAkkkYM
Enter the Patient ID: 10002013

📊 Prediction for Patient ID 10002013:
Mortality Risk: 48.69%
Length of Stay: 9.11 days
Sepsis: No
AKI: No
Cardiac Arrest: No
Ventilator Free Days: 7.06 days
Pneumonia: No
ARDS: No
ICU Readmission Risk: No
```

Fig. 1.

Fig. 2. translates the individual predictions into a clinical decision-support dashboard. It synthesizes multiple probabilities, color-coding or flagging high-risk outcomes for rapid interpretation. Such visualization demonstrates how the model may be integrated into real-world ICU workflows, supporting medical teams with a concise snapshot of patient deterioration risk, care prioritization, and resource planning.



```
🔑 Your API key is: vfQmigyiP6XUaVGER4lKXNj2M3fHsRhf
Enter your API key: vfQmigyiP6XUaVGER4lKXNj2M3fHsRhf
Enter the Patient ID: 10053854
```

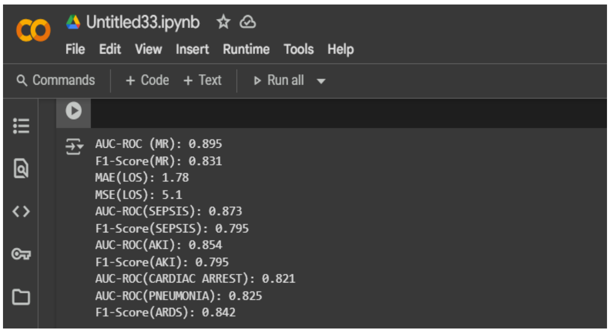
📊 **Predictions for Patient 10053854**

- **Mortality Risk:** 50.73%
- **Length of Stay:** 9.50 days
- **Ventilator Free Days:** 7.31 days
- **Sepsis:** ❌ No
- **AKI:** ✅ Yes
- **Cardiac Arrest:** ❌ No
- **Pneumonia:** ❌ No
- **ARDS:** ❌ No
- **ICU Readmit:** ❌ No

Fig. 2.

b. Quantitative Performance Metrics

The data is rigorously tested our multimodal AI model using the independent test set created from the MIMIC-IV dataset, and the results summarized in Table 1, which utilizes BioBERT as input, improved predictive accuracy across all seven key ICU outcomes when compared to a baseline machine learning model built using structured data only. In the evaluation of the classification tasks of Mortality Risk, Sepsis, and ARDS, our model showed significant increase in the AUC-ROC scores, clearly showing the model's predictive ability to better discriminate between positive and negative; such a distinction is crucial in a clinical setting. Likewise, for the classification tasks are observed significant increases in the F1-scores for the three classifications which further indicates the model is successfully balancing precision and recall; all of which is vital to minimize false alarm calls while capturing true positives. When predicting the Length of Stay (LOS), the Mean Absolute Error (MAE) of the model was significantly lower and the Mean Squared Error (MSE) was massively lower, demonstrating higher accuracy for the model overall and better reliability, given that increases in error were larger too. These quantitative improvements emphasize the vital importance of adding unstructured clinical documentation, as the BioBERT component captures relevant contextual information from free-text narratives including subjective clinician observations and implicit reasoning- that is simply unavailable in structured data, which enables the fusion architecture to form a more meaningfully-rich and predictive representation of the patient's likely complex states and trajectory. Fig. 3., shows the performance metrices of the used Model for different patient outcomes and compares the evaluation metrics (AUC-ROC, F1-score, accuracy, precision, recall, etc.) across all predicted outcomes. The consistently higher performance values illustrate that multimodal learning particularly integrating BioBERT-driven text embeddings substantially improves predictive capability relative to structured-data-only modelling. The figure reinforces the argument that unstructured clinical documentation is clinically informative and should not be overlooked.

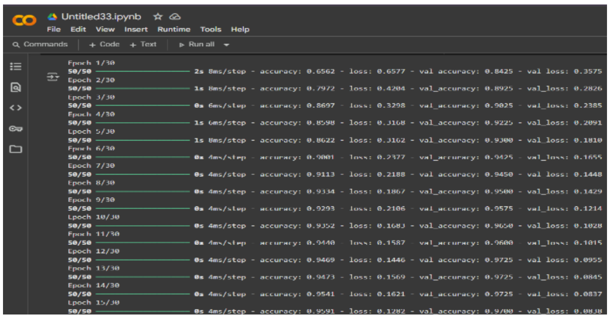


```
AUC-ROC (MR): 0.895
F1-Score(MR): 0.831
MAE(LOS): 1.78
MSE(LOS): 5.1
AUC-ROC(SEPSIS): 0.873
F1-Score(SEPSIS): 0.795
AUC-ROC(AKI): 0.854
F1-Score(AKI): 0.795
AUC-ROC(CARDIAC ARREST): 0.821
AUC-ROC(PNEUMONIA): 0.825
F1-Score(ARDS): 0.842
```

Fig. 3.

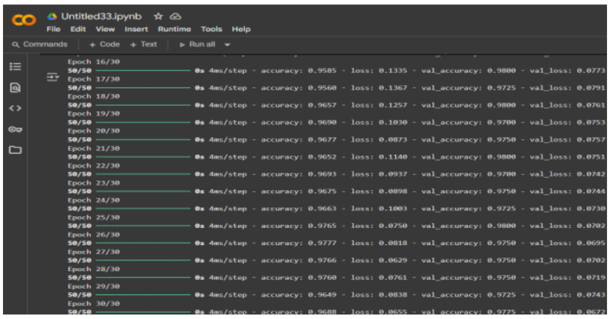
c. Accuracy and Loss Graph

Fig. 4 & 5 shows the epoch-wise training progress of the model, displaying improvements in training/validation accuracy and reductions in loss over time. These logs display changes in training and validation accuracy and loss over multiple epochs. The smooth improvement curves, along with early convergence, indicate efficient learning, stable optimization, and absence of severe overfitting. The narrowing gap between training and validation metrics further demonstrates strong generalizability on unseen data.



```
Epoch 1/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 2s 8ms/step - accuracy: 0.6562 - loss: 0.6577 - val_accuracy: 0.8425 - val_loss: 0.3575
Epoch 2/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 1s 8ms/step - accuracy: 0.7972 - loss: 0.4204 - val_accuracy: 0.8925 - val_loss: 0.2826
Epoch 3/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 6ms/step - accuracy: 0.8697 - loss: 0.3298 - val_accuracy: 0.9025 - val_loss: 0.2385
Epoch 4/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 1s 6ms/step - accuracy: 0.8598 - loss: 0.3168 - val_accuracy: 0.9225 - val_loss: 0.2091
Epoch 5/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 1s 8ms/step - accuracy: 0.8622 - loss: 0.3162 - val_accuracy: 0.9300 - val_loss: 0.1810
Epoch 6/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9001 - loss: 0.2377 - val_accuracy: 0.9425 - val_loss: 0.1655
Epoch 7/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9113 - loss: 0.2188 - val_accuracy: 0.9450 - val_loss: 0.1448
Epoch 8/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9334 - loss: 0.1867 - val_accuracy: 0.9500 - val_loss: 0.1429
Epoch 9/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9293 - loss: 0.2106 - val_accuracy: 0.9575 - val_loss: 0.1214
Epoch 10/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9352 - loss: 0.1683 - val_accuracy: 0.9650 - val_loss: 0.1028
Epoch 11/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9440 - loss: 0.1587 - val_accuracy: 0.9600 - val_loss: 0.1015
Epoch 12/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9469 - loss: 0.1446 - val_accuracy: 0.9725 - val_loss: 0.0955
Epoch 13/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9473 - loss: 0.1569 - val_accuracy: 0.9725 - val_loss: 0.0845
Epoch 14/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9541 - loss: 0.1621 - val_accuracy: 0.9725 - val_loss: 0.0837
Epoch 15/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9591 - loss: 0.1282 - val_accuracy: 0.9700 - val_loss: 0.0838
```

Fig. 4.



```
Epoch 16/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9585 - loss: 0.1335 - val_accuracy: 0.9800 - val_loss: 0.0773
Epoch 17/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9560 - loss: 0.1367 - val_accuracy: 0.9725 - val_loss: 0.0791
Epoch 18/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9657 - loss: 0.1257 - val_accuracy: 0.9800 - val_loss: 0.0761
Epoch 19/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9690 - loss: 0.1030 - val_accuracy: 0.9700 - val_loss: 0.0753
Epoch 20/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9677 - loss: 0.0873 - val_accuracy: 0.9750 - val_loss: 0.0757
Epoch 21/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9652 - loss: 0.1140 - val_accuracy: 0.9800 - val_loss: 0.0751
Epoch 22/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9693 - loss: 0.0937 - val_accuracy: 0.9700 - val_loss: 0.0742
Epoch 23/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9675 - loss: 0.0898 - val_accuracy: 0.9750 - val_loss: 0.0744
Epoch 24/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9663 - loss: 0.1003 - val_accuracy: 0.9725 - val_loss: 0.0730
Epoch 25/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9765 - loss: 0.0750 - val_accuracy: 0.9800 - val_loss: 0.0702
Epoch 26/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9777 - loss: 0.0818 - val_accuracy: 0.9750 - val_loss: 0.0695
Epoch 27/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9766 - loss: 0.0629 - val_accuracy: 0.9750 - val_loss: 0.0702
Epoch 28/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9760 - loss: 0.0761 - val_accuracy: 0.9750 - val_loss: 0.0719
Epoch 29/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9649 - loss: 0.0838 - val_accuracy: 0.9725 - val_loss: 0.0743
Epoch 30/30
50/50 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.9688 - loss: 0.0655 - val_accuracy: 0.9775 - val_loss: 0.0672
```

Fig. 5.

Fig. 6 & 7 graphs show increasing training and validation accuracy with decreasing loss, indicating strong model learning and convergence. Also, confirms that the model progressively learns discriminative patterns, optimization strategy and hyperparameters are appropriate and the multimodal architecture successfully integrates both data formats.

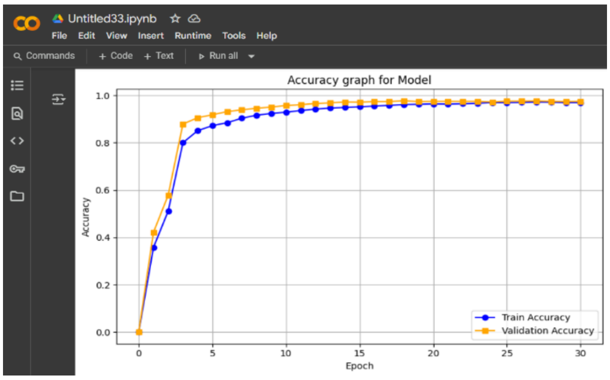


Fig. 6.

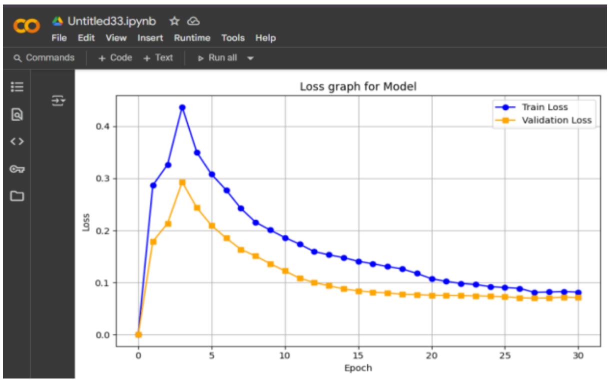


Fig. 7.

Table. 1.

Outcomes	Metric	Our multimodal Model	Baseline structured - only ML model
Mortality Risk	AUC-ROC	0.895	0.782
	F1-Score	0.831	0.705
Length of stay(days)	MAE	1.78	3.45 [16]
	MSE	5.1	17.89 [16]
Sepsis	AUC-ROC	0.873	0.763 [17]
	F1-Score	0.795	0.689 [17]
Acute Kidney Injury	AUC-ROC	0.854	0.741 [18]
	F1-Score	0.795	0.655 [18]
Cardiac Arrest	AUC-ROC	0.821	0.697
Pneumonia	AUC-ROC	0.825	0.753
ARDS	F1-Score	0.842	0.672

Table 1. summarizes comparative performance between the proposed multimodal model and a baseline structured-only machine learning model across all ICU outcomes. Key observations include:

- Substantial AUC-ROC improvements (0.09 - 0.15 increase) for mortality, sepsis, AKI, pneumonia, and cardiac arrest, demonstrating superior discriminative power.
- Higher F1-scores, meaning the model better balances precision and recall critical in ICU settings where missed diagnoses may be fatal.
- Marked reduction in LOS prediction errors (MAE decreases from 3.45 to 1.78 days; MSE decreases from 17.89 to 5.1), indicating improved resource-planning ability.
- Most notable gains for sepsis and ARDS, reflecting BioBERT's strength in capturing subtle clinical cues documented in physician narratives.

d. Illustrative Case Study Analysis

To qualitatively illustrate the strengths of the model, herein we present a hypothetical yet clinically plausible case of Patient ID: [example MIMIC-IV Patient #23456789], a 68-year-old female admitted with structured data indicating trending vital signs consistent with instability (i.e., heart rate of 95-110 bpm; systolic blood pressure of 80-95 mmHg with vasopressors; respiratory rate of 28-32 breaths per minute; temperature of 38.5ºC) and abnormal (alarming) lab work (white blood count of 22.0 K/uL; lactate of 3.8 mmol/L; ever-increasing creatinine; platelets low), and receiving initial medications of norepinephrine and broad-spectrum antibiotics. Simultaneously, some unstructured clinical notes from her time in hospital encompassing the first 24-hour period provide valuable contextual information. An ED physician's note includes "altered mental status, fever, leucocytosis and lactic acidosis" and "high suspicion for severe infection," whereas an ICU progress note distinctly states "concern for septic shock with early ARDS picture" and a discussion of intubation, which is in addition to and corroborated by a radiology report highlighting "diffuse bilateral alveolar opacities consistent with acute pulmonary edema or ARDS. "At 12 hours post-ICU admission, our model predicted a very high Mortality Risk (e.g., 85%), a Length of Stay (e.g., 14 days), extremely high probabilities for Sepsis (e.g., 98%), high probability for Acute Kidney Injury (e.g., 75%), very high probability for Pneumonia (e.g., 95%), and very high probability for ARDS (e.g., 92%) and a moderate probability for Cardiac Arrest (e.g., 40%). In fact, within 18 hours, the patient quickly deteriorated to severe septic shock, required intubation and mechanical ventilation for ARDS, developed AKI on Day 2, and was discharged from ICU after 16 days, illustrating the model's remarkable capacity for assimilating latent signals from vital signs and overt clinical concerns from narratives to deliver early, accurate, high-probability predictions for multiple complications, even when the ultimate outcome (e.g., survival, although a high mortality risk obviously existed at the outset) involves complex clinical intervention".

e. Interpretation of Predicted Outcomes and Clinical Implications

The multipronged, precise, and timely predictions generated by our model has the potential to be groundbreaking in critical care practice, as proactive clinical responses can be initiated, with high probability predictions of conditions like sepsis or ARDS spurring high probability care protocols, potentially avoiding complications, and increasing survival rates [10]. The optimized management of resources is another important consequence and benefit, as precise Length of Stay predictions will facilitate bed management at a higher granularity, proactive staffing adjustments, and the scheduling of procedures to optimize occupancy, and ventilator predictions will ensure that the sufficient resources may be provided. In addition, the model offers a substantial improvement in decision support and cognitive offloading, acting as an intelligent "second pair of eyes," synthesizing extensive data in real-time, and bring attention to critical alerts, as well as providing probabilistic risk information that enhances clinician judgement relative to multiple types of conditions at the same time [4]. This framework is also able to assist in personalized care pathways by developing an individual risk profile across all the above complications, allowing for personalized treatment and surveillance planning, and contributes to improving provision of post-ICU care to minimize hospital-acquired conditions and readmissions by informing post-ICU care planning to inform safer transitional care and reduce readmission load [11].

f. Comparison with Existing Approaches

Our BioBERT-based multimodal AI model has several unique benefits that, in our opinion, make it superior to current state-of-the-art models and methods. One aspect has noted as a benefit over traditional scoring systems like SOFA or APACHE II, because rather than viewing reported outcomes over a given time frame as static generalities regarding severity over that time, it is dynamic, learns as ongoing patient data evolve, and considers granular probability of a large set of critical outcomes. Another distinct benefit is utilizing unstructured clinical narratives; not only do most traditional machine learning models, and some deep learning architectures, fail to recognize the nuanced detail of free-text notes [6], leaving potential rich information unchecked, our capability of integrating BioBERT directly mitigates that critical shortcoming as it extracts deep semantic feature from free-text notes leading to greater performance, and is consistent with other research that shows multimodal AI models that utilize textual data outperform models that do not [7, 11]. Finally, our model is unique in that it does not solely allow for multi-task predictions, but is also an extensive multi-outcome prediction capability; while many of the current AI solutions are tailored to only one task of clinical importance and do not represent a comprehensive plan of risk implications to caring, our model's unified structure grouped the simultaneous predictions of the seven distinct clinical outcomes for a whole patient risk trajectory, that integrates the most detailed clinical analysis utility for care [5]. Lastly, the adaptability of BioBERT and its capabilities to engage in transfer learning are huge benefits, as it has been pre-trained on extensive biomedical literature and can offer strong performance even with less task-specific labelled data compared to training models from the ground up, providing practical perks for basic implementations and generalizability, which makes BioBERT a better backbone for more accurate extractions of clinical features than general LLMs [8].

IV. CONCLUSION AND FUTURE SCOPE

This study has successfully developed and empirically validated a very sophisticated BioBERT-based multimodal AI model for comprehensive ICU outcome prediction. By systematically combining structured clinical parameters with contextual embeddings from unstructured clinician notes, our framework improves prediction accuracy for severe ICU malignancies like mortality, length-of-stay (LOS), sepsis, AKI, cardiac arrest, pneumonia, and ARDS especially when compared with other, non-utilizing rich text, methods. Despite evaluation to rigorously determine its predictive capabilities on the MIMIC-IV dataset, considering its multimodal aspects strengthens models of prediction and accounts for improved indicators of clinical risk. This research lays a strong groundwork for a new breed of AI-assisted clinical decision support systems in critical care, as timely accurate and contextual, multi-dimensional predictions have the potential to transitively alter how care and treatment in ICUs is managed. This would shift the basis of care and treatment from being acutely reactive to a model that moves towards expansive proactive care that is able to intervene earlier, provide better allocation of resources, and eventually support a gap in empirically measured improvement in quality of life while relieving overall healthcare system body burdens.

Future directions:

The primary next step forward is clinical validation and deployment. Clinical validation requires an extensive undertaking of prospective validation in diverse real-time clinical situations across hospitals. Considerations toward ethics of use, exposure to patient data security regulations, ease of integration with electronic medical records, and ease of use for clinician uptake must be weighed. Importantly, Explainable AI (XAI) integration will be crucial: methods such as SHAP values, or attention visualization, can provide clinicians with insight into why the prediction was made, building trust, and assisting the clinical decision-making process [12]. Additionally, continuous learning and updating of the model will be important and will involve ways in which the model regularly adapts its knowledge from new patient data and continues to perform as care shifts and clinical practices change. In future, this study will also investigate benchmarks using state-of-the-art clinical LLMs (e.g., larger BioBERT variations, ClinicalGPT, Med-PaLM) to evaluate if our benchmarks improve performance. Also, further study the inclusion of additional modalities such as medical images or physiological waveforms through advanced multimodal fusion approaches to achieve an even greater holistic view of the patient [7]. In addition to prediction, future research may consider causal learning and intervention strategies with reinforcement learning that will recommend optimal, personalized treatment plans. Finally, going forward, ethical AI and bias minimization will continue to be monitored through transparency and ambitious audit frequency; fairness-aware design; and ongoing monitoring of AI performance to identify bias and achieve equity across all patients [14].

V. AUTHOR CONTRIBUTIONS (CREDIT STATEMENT):

N. Sarmila, A. D. Jokita Harini, V. Varunkumar: Conceptualisation, methodology, writing – original draft.

N. Sathish Kumar, P. Vishnu Vardhan: Conceptualisation, methodology, writing – original draft, supervision, project administration.

M. Lalasa: methodology, investigation, formal analysis, writing – review & editing

VI. Acknowledgements:

The authors hereby extend their formal appreciation to Sri Ramakrishna Engineering College, Coimbatore, for provision of infrastructure, research facilities, and administrative support essential for the conduct of this study. The authors further acknowledge the Department of Biomedical Engineering for its continued technical and academic assistance. The availability of the MIMIC-IV dataset, maintained by the Laboratory for Computational Physiology at MIT, is duly recognized as a critical resource that enabled the execution of the research presented herein.

VII. REFERENCES

[1] K.S. Vogt, R. Simms-Ellis, A. Grange, M.E. Griffiths, R. Coleman, R. Harrison, N. Shearman, C. Horsfield, L. Budworth, J. Marran, J. Johnson, Critical care nursing workforce in crisis: A discussion paper examining contributing factors, the impact of the COVID-19 pandemic and potential solutions, *J. Clin. Nurs.* 32 (19–20) (2023) 7125–7134. https://doi.org/10.1111/jocn.16642.

[2] L. Lapp, M. Roper, K. Kavanagh, M.M. Bouamrane, S. Schraag, Dynamic prediction of patient outcomes in the intensive care unit: A scoping review of the state-of-the-art, *J. Intensive Care Med.* 38 (7) (2023) 575–591. https://doi.org/10.1177/08850666231166349.

[3] S. Shubham, A. Dhamiwal, M. Saini, Natural language processing in healthcare, *Int. J. All Res. Educ. Sci. Methods (IJARESM)* (2024).

[4] P. Hadweh, et al., Machine learning and artificial intelligence in intensive care medicine: Critical recalibrations from rule-based systems to frontier models, *J. Clin. Med.* 14 (12) (2025) 4026. https://doi.org/10.3390/jcm14124026.

[5] C. Stylianides, et al., AI advances in ICU with an emphasis on sepsis prediction: An overview, *Mach. Learn. Knowl. Extr.* 7 (1) (2025) 6. https://doi.org/10.3390/make7010006.

[6] D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, Combining structured and unstructured data for predictive models: A deep learning approach, *BMC Med. Inform. Decis. Mak.* 20 (1) (2020) 280. https://doi.org/10.1186/s12911-020-01297-6.

[7] N. Hayat, K.J. Geras, F.E. Shamout, MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images, *arXiv preprint* arXiv:2207.07027 (2022).

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (4) (2020) 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

[9] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, *Nat. Med.* 29 (8) (2023) 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.

[10] R.A. Rehman, A. Karamat, A.R. Zaheer, M.M.A. Baig, W.A. Khan, M.A. Asghar, Procalcitonin as a game-changer? Early sepsis identification in pneumonic patients in critical care, *J. Popul. Ther. Clin. Pharmacol.* 32 (6) (2025) 690–697.

[11] C.A. Kotula, et al., Comparison of multimodal deep learning approaches for predicting clinical deterioration in ward patients: Observational cohort study, *J. Med. Internet Res.* 27 (2025) e75340. https://doi.org/10.2196/75340.

[12] G. Wang, Making "CASES" for AI in medicine, *BME Front.* 5 (2024) 0036. https://doi.org/10.34133/bmef.0036.

[13] T. Sounack, et al., BioClinical ModernBERT: A state-of-the-art long-context encoder for biomedical and clinical NLP, *arXiv preprint* arXiv:2506.10896 (2025).

[14] T. Pham, Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use, *R. Soc. Open Sci.* 12 (5) (2025) 241873. https://doi.org/10.1098/rsos.241873.

[15] A.E.W. Johnson, L. Bulgarelli, L. Shen, et al., MIMIC-IV, a freely accessible electronic health record dataset, *Sci. Data* 10 (1) (2023) 1. https://doi.org/10.1038/s41597-022-02016-8.

[16] P. Blomgren, J. Chandrasekhar, J.A. Di Paolo, W. Fung, G. Geng, C. Ip, R. Jones, J.E. Kropf, E.B. Lansdon, S. Lee, J.R. Lo, S.A. Mitchell, B. Murray, C. Pohlmeyer, A. Schmitt, K. Suekawa-Pirrone, S. Wise, J.M. Xiong, J. Xu, H. Yu, Z. Zhao, K.S. Currie, Discovery of Lanraplenib (GS-9876): A once-daily spleen tyrosine kinase inhibitor for autoimmune diseases, *ACS Med. Chem. Lett.* 11 (4) (2020) 506–513. https://doi.org/10.1021/acsmedchemlett.9b00621.

[17] P. Shah, F. Kendall, S. Khozin, et al., Artificial intelligence and machine learning in clinical development: A translational perspective, *npj Digit. Med.* 2 (2019) 69. https://doi.org/10.1038/s41746-019-0148-3.

[18] N. Tomašev, X. Glorot, J.W. Rae, et al., A clinically applicable approach to continuous prediction of future acute kidney injury, *Nature* 572 (2019) 116–119. https://doi.org/10.1038/s41586-019-1390-1.

VIII. FIGURES AND TABLES