Advanced Multimodal AI for Resilient Healthcare: Enhancing Early Risk Assessment in Critical Care

Shih-Wei Wu[1], Chengcheng Li[2], Te-Nien Chien[3]*, Yao-Yu Zhang[4]
Department of Business Management, National Taipei University of Technology, Taipei, Taiwan [1]
College of Management, National Taipei University of Technology, Taipei, Taiwan [2, 3]
Division of Cardiology-Department of Internal Medicine, Taipei Veterans General Hospital, Taipei, Taiwan [4]

***Abstract*—This study develops an advanced multimodal AI framework to strengthen early risk assessment in critical care and support resilient healthcare delivery. Utilizing the MIMIC-III database, this research extracted structured variables and clinical notes from 26,829 adult patients. A text mining approach based on the BERTopic model was employed to generate topic embeddings from unstructured notes, which were subsequently integrated with 16 quantitative variables. Six machine learning models: Adaboost, Gradient Boosting, Support Vector Classification (SVC), Bagging, Logistic Regression, and MLP Classifier were trained to predict short-term and long-term mortality outcomes. Model performance was evaluated through AUROC, accuracy, recall, precision, and F1-score metrics. The results demonstrate that integrating topic embeddings with structured data significantly improved short-term risk prediction. The SVC model, in particular, achieved an AUROC of 0.9137 for predicting 2-day mortality. Critical predictors identified included the Glasgow Coma Scale, White Blood Cell Count, and text-derived topics related to cardiovascular and neurological conditions. The study is based on a single-center dataset, limiting generalizability. Additionally, only a subset of textual data sources was analyzed, and improvements in long-term risk prediction were relatively modest. These findings demonstrate how multimodal AI can significantly improve early risk assessment and enhance resilience in critical care decision-making. This research pioneers the integration of BERTopic-based text mining with machine learning models for clinical risk prediction, highlighting the value of multimodal data fusion in improving predictive accuracy and enriching medical informatics.**

Keywords*—*Resilient healthcare; multimodal AI; early risk assessment; critical care; clinical text mining

I. Introduction

Artificial Intelligence (AI) and healthcare big data are revolutionizing clinical outcome prediction and decision-making, particularly in critical care environments such as Intensive Care Units (ICUs) [1]. AI, powered by sophisticated machine learning algorithms, processes extensive and complex datasets to uncover intricate patterns and generate precise predictions of patient outcomes. These advancements significantly enhance clinical decision-making processes and enable the delivery of highly personalized care. Within this framework, Electronic Health Records (EHRs) play a pivotal role as the foundation of healthcare big data. EHRs encompass comprehensive digital records of patients' health information, including medical histories, diagnostic imaging, and treatment details [2]. The digital transformation of health records not only streamlines hospital management and service delivery but also provides researchers with vast datasets for developing and validating predictive models [3]. These technological advancements further enhance predictive analytics and play a critical role in supporting resilient healthcare by enabling earlier identification of clinical deterioration and improving system responsiveness [4, 5].

In addition to numerical and structured data in EHRs, semi-structured and unstructured data, such as clinical notes, diagnostic reports, and patient discharge summaries, constitute a valuable yet often underutilized source of information in predictive modeling [6]. Text mining, a specialized area within AI dedicated to extracting meaningful insights from unstructured text, has gained significant recognition for its ability to process such data. By leveraging advanced text mining techniques, researchers can uncover critical insights from clinical notes and reports, transforming free-text into structured data that can seamlessly integrate with traditional numerical datasets to enhance predictive modeling accuracy.

The integration of text mining with healthcare big data comprising EHRs, diagnostic reports, and other patient-related records provides a powerful foundation for developing predictive models to support evidence-based clinical decision-making [7]. This synergy holds considerable promise for enhancing resource allocation, advancing patient care, and minimizing healthcare expenditures, establishing it as a central focus in contemporary medical informatics research. In ICU-focused studies, leveraging EHR data for predictive tasks, such as assessing patient mortality, estimating length of stay, and diagnosing diseases, plays a pivotal role. Effectively utilizing EHR data for clinical outcome prediction enables early identification of high-risk patients, facilitates timely interventions, and improves mortality risk assessment, ultimately optimizing patient care and resource utilization [8]. Strengthening early risk prediction in the ICU is therefore fundamental to resilient healthcare, as it supports timely interventions, reduces avoidable deterioration, and enhances the system's capacity to respond effectively under clinical pressure [9-11].

Accurate prediction of clinical outcomes for ICU patients, given their diverse survival probabilities, plays a crucial role in enhancing the quality of care within ICUs, facilitating cross-institutional evaluations, and advancing related clinical research [12]. Numerous predictive methods and scoring systems, including APACHE, SAPS, and MPM, have been designed to estimate clinical outcomes [13, 14]. While these models achieve

*Corresponding author.

reasonable accuracy in predicting ICU mortality rates, they often require recalibration when applied to different ICU settings due to disparities in patient demographics and clinical practices [15]. Despite these advancements, there remains a significant knowledge gap in how advanced topic modeling like BERTopic can be effectively integrated with structured ICU data to improve predictive resilience. This study fills this gap by providing a comparative analysis of multimodal AI frameworks. Machine learning (ML) has emerged as a transformative tool for data analysis, offering robust capabilities in feature extraction and addressing the limitations of traditional statistical approaches.ML techniques have been widely implemented in various healthcare predictive tasks, such as employing Random Forest algorithms to assess fetal maturity and adapt model parameters across diverse temporal settings [16].

While EHR adoption expands, research remains predominantly focused on structured quantitative data. Yet, clinical narratives containing vital documentation of physiological states and disease progression represent nearly 80% of EHR content [17, 18]. These semi-structured records encode essential clinical knowledge, including diagnostic assessments, treatment protocols, and care priorities, which are critical for evidence-based decision-making [19]. Natural language processing (NLP) has consequently become indispensable for deriving insights from such textual data [20, 21], exemplified by its 90% accuracy in classifying peripheral arterial disease through keyword extraction [22].

While NLP improves patient classification accuracy, it simultaneously reveals challenges, including data sparsity and high dimensionality. To address these limitations, we propose the BERTopic model, which integrates BERT (Bidirectional Encoder Representations from Transformers) embeddings with c-TF-IDF to generate dense clusters and extract interpretable topics. In contrast to conventional latent Dirichlet allocation (LDA) approaches, BERTopic provides enhanced semantic representations that resolve vocabulary mismatches while capturing temporal dynamics in topic distributions [23, 24]. Unlike LDA, which relies on bag-of-words, BERTopic utilizes contextual embeddings to maintain semantic coherence even in short, noisy clinical notes. Furthermore, the model streamlines hyperparameter optimization and incorporates dedicated noise topic handling, effectively reducing misclassification of irrelevant documents and consequently improving overall topic modeling accuracy [25, 26].

Integrating text mining methods like BERTopic with machine learning provides a robust framework for enhancing predictive accuracy in critical care. While conventional models often overlook the rich clinical insights within semi-structured data, such as nursing notes and diagnostic reports, this study leverages the BERTopic model to extract and categorize meaningful latent topics. By synthesizing these text-derived features with numerical datasets, our approach enables a more holistic understanding of patient conditions. This integration transforms unstructured narratives into quantifiable predictors, equipping healthcare professionals with deeper insights for more precise decision-making and ICU outcome forecasting. This study advances the field by employing a transformer-enhanced BERTopic model to automatically extract interpretable features from complex clinical narratives. A key contribution of this work is the comprehensive comparative analysis that provides empirical evidence of the advantages modern topic modeling holds over traditional methods. As healthcare big data analytics evolve, incorporating advanced text mining into predictive frameworks promises to optimize resource allocation and enhance the quality of patient care. Ultimately, this multimodal AI framework reinforces early risk recognition, supporting resilient healthcare by enabling more timely, adaptive, and effective interventions in critical care settings.

II. Materials and Methods

The methodological framework of this study is illustrated in Fig. 1. The dataset consists of structured EHR variables and clinical notes, providing a comprehensive representation of patient information. Data extraction, cleaning, and preprocessing were performed to ensure analytical quality. The processed dataset was then used to train five machine learning models for mortality prediction. Model performance was evaluated using five standard metrics to enable an objective comparison. The following sections describe each methodological component in detail.

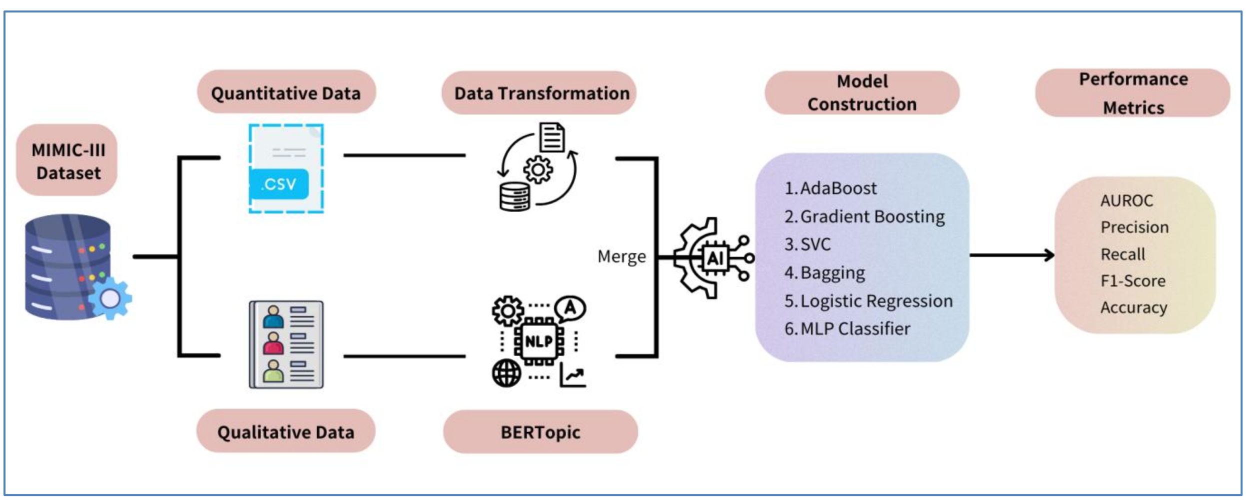

Fig. 1. Research scheme.

A. Data Composition and Source: MIMIC-III

This study utilized data from the Medical Information Mart for Intensive Care III (MIMIC-III), a publicly available clinical database that contains detailed records of intensive care unit (ICU) patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts. The dataset encompasses patient encounters from 2001 to 2012, capturing a wide range of clinical variables, including vital signs, medication history, laboratory test results, and observational notes [27]. MIMIC-III includes 49,785 hospital admissions, representing 38,597 unique adult patients aged 16 years and older, with a median age of 65.8 years and a male patient proportion of 55.9%.

To ensure ethical compliance, access to the MIMIC-III database required approval under certificate number 35628530, which involved completing the National Institutes of Health (NIH) online training, successfully passing the Human Research Participant Protection Examination, and submitting a formal data access request. Additionally, the study obtained institutional review board (IRB) approval from both BIDMC and the Massachusetts Institute of Technology (MIT). The MIMIC-III dataset is curated by the MIT Laboratory for Computational Physiology and is available for research purposes through PhysioNet. This rigorous approval process ensures adherence to ethical guidelines and data protection regulations, enabling responsible use of de-identified patient data for advancing critical care research. (https://physionet.org/content/mimiciii/1.4/)

B. Data Preprocessing

1) Data extraction: To enhance the generalizability of the findings, this study encompassed all ICU patients rather than limiting the analysis to specific disease groups. For consistency with similar research, adult ICU patients aged over 16 years and admitted to the ICU for the first time were included. The analysis focused primarily on data recorded during the initial 24 hours of these patients' ICU stays [28, 29]. Table I summarizes the demographic and clinical characteristics of the patient cohort following data preprocessing. The final cohort comprised 26,829 patients who met the inclusion criteria and had associated clinical notes available for textual analysis, including 2,322 (8.66%) who died in the hospital and 24,507 (91.34%) who survived. Admission Type refers to the classification of the patient's admission leading to the index ICU stay. The median patient age was 63.06 years, with 56.85% of the cohort being male. The majority of patients were identified as white (71.20%), followed by black (7.70%) and individuals from other ethnic groups. Emergency admissions accounted for 82.29% of cases, with 37.15% of patients admitted to Medical ICUs. The average length of stay in the ICU was 4.15 days.

2) Variable selection: To identify the variables for analysis, this study leveraged insights from prior research and selected 16 quantitative variables based on their clinical significance [listed in Appendix A (Table VIII)] [30-32]. These variables were extracted from multiple tables within the MIMIC-III dataset, including admission records, chartevents, labevents, and output events. A structured three-stage data preprocessing approach, adapted from Guo et al. [8] was implemented to address missing values. Initially, patient records with more than 30% missing data were excluded. Next, predictors with over 40% missing values were removed. Finally, variables with a missing rate exceeding 20% after applying the previous filtering steps (i.e., exclusion of patient records with >30% missing data and exclusion of predictors with >40% missing values) were eliminated. For the remaining missing values, mean imputation was performed. The significance of the selected variables was evaluated using the Information Gain Technique (Entropy) [33], retaining only those with a score of 0.01 or higher for further analysis. Among the identified predictors, white blood cell count emerged as the most influential, whereas gender ranked the lowest.In addition to numerical variables, this study incorporated topic modeling-derived features from the NOTEVENTS dataset, which comprises clinical notes documented by various healthcare professionals, including physicians, nurses, imaging specialists, nutritionists, and physical therapists. Within the MIMIC-III database, NOTEVENTS contains over 2 million entries, with approximately 56% authored by doctors and nurses. An additional 39% of the records consist of echocardiography, electrocardiography, and radiology reports. By integrating structured numerical data with unstructured textual information, this study aimed to enhance predictive accuracy and provide a more comprehensive assessment of ICU patient outcomes. This multimodal approach enables a more robust representation of clinical contexts that would otherwise be overlooked when relying solely on structured data.

C. Mortality Prediction

This study examined the impact of integrating structured data (e.g., vital signs and laboratory test results) with semi-structured data (e.g., diagnostic details and clinical notes) on predicting ICU patient mortality following admission. This study aimed to predict ICU patient mortality at specific time points following the initial ICU admission, using data collected within the first 24 hours. We focused on four distinct prediction targets:

1) Short-term mortality: Defined as patient death occurring within 2 days or 3 days post-ICU admission.

2) Long-term mortality: Defined as patient death occurring within 1 month or 1 year post-ICU admission.

The integration of these data types aimed to improve the precision and reliability of mortality predictions across both short-term and long-term timeframes.

D. BERTopic

BERTopic represents a cutting-edge approach to topic modeling, harnessing the strengths of BERT (Bidirectional Encoder Representations from Transformers) to uncover latent themes within large textual datasets [34]. This method excels in generating coherent topic representations by employing a combination of BERT-based embeddings and a class-based variation of TF-IDF (c-TF-IDF). The BERTopic algorithm consists of three key stages: document embedding, dimensionality reduction using UMAP, and document clustering with HDBSCAN.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF SELECTED PATIENTS

	Overall	Dead at Hospital	Alive at Hospital
General (%)			
Number	26,829 (100%)	2,322 (8.66%)	24,507 (91.34%)
Age [Q1–Q3]	63.06 [51.32–77.82]	70.88 [61.43–83.32]	62.26 [50.51–76.97]
Gender (male)	15,248 (56.85%)	1,256 (8.24%)	13,992 (91.76%)
Race/Ethnicity (%)			
Asian	656 (2.45%)	57 (2.45%)	599 (2.44%)
Black	2,067 (7.70%)	109 (4.69%)	1,958 (7.99%)
Hispanic	888 (3.31%)	47 (2.02%)	841 (3.43%)
White	19,105 (71.20%)	1,673 (72.05%)	17,432 (71.11%)
Other	4,113 (15.34%)	601 (25.88%)	3,512 (14.33%)
Admission Type (%)			
Urgent	644 (2.40%)	74 (3.19%)	570 (2.33%)
Emergency	22,067 (82.29%)	2,199 (94.70%)	19,868 (81.07%)
Elective	4,118 (15.32%)	49 (2.11%)	4,069 (16.60%)
Site (%)			
Medical Intensive Care Unit	9,966 (37.15%)	1,125 (48.45%)	8,841 (36.06%)
Surgical Intensive Care Unit	4,400 (16.40%)	463 (19.94%)	3,937 (16.06%)
Coronary Care Unit	4,180 (15.58%)	326 (14.04%)	3,854 (15.73%)
Cardiac Surgery Recovery Unit	4,336 (16.16%)	121 (5.21%)	4,215 (17.20%)
Trauma Surgical Intensive Care Unit	3,947 (14.71%)	287 (12.36%)	3,660 (14.93%)
Insurance			
Government	819 (3.05%)	45 (1.94%)	774 (3.16%)
Medicaid	2,235 (8.33%)	144 (6.20%)	2,091 (8.53%)
Medicare	14,342 (53.46%)	1,621 (69.81%)	12,721 (51.91%)
Private	9,080 (33.84%)	459 (19.76%)	8,621 (35.18%)
Self-Pay	353 (1.32%)	53 (2.28%)	300 (1.22%)
Outcomes			
Hospital LOS (days) [Q1–Q3]	8.95 [3.88–10.47]	9.27 [2.77–11.49]	8.92 [3.96–10.34]
ICU LOS (days) [Q1–Q3]	4.15 [1.26–4.17]	6.68 [2.08–8.12]	3.89 [1.22–3.89]

1) Bert embedding: In the first stage of the BERTopic algorithm, BERT Embeddings are used to convert documents into dense vector representations, capturing the semantic relationships within the text. BERT (Bidirectional Encoder Representations from Transformers), a transformer-based model, is pre-trained on large corpora using masked language modeling, allowing it to consider both the left and right context of words in a sentence. Sentence-BERT (SBERT), a variation of BERT, is employed to create sentence-level embeddings optimized for semantic similarity. During this process, each document is tokenized into subwords using BERT's Word Piece tokenizer, ensuring that rare or unseen words are broken down into smaller, meaningful units. The tokenized sequence is then passed through BERT's multi-layer attention mechanism to generate contextual embeddings for each token. These token embeddings are aggregated, typically using the embedding of the [CLS] token, which represents the entire document. The final document embedding, D, can be expressed as:

$$D = f_{BERT}(T) = h_{[CLS]} \quad (1)$$

where, T is the tokenized input sequence and $h_{[CLS]}$ is the hidden state corresponding to the [CLS] token, which summarizes the semantic content of the entire document. The document embedding process plays a critical role by allowing the algorithm to place semantically similar documents in close proximity within the vector space. This step establishes the foundation for subsequent stages, including clustering and topic extraction [34].

2) Uniform manifold approximation and projection (UMAP): In the second stage of the BERTopic algorithm, Uniform Manifold Approximation and Projection (UMAP) is employed to reduce the dimensionality of the high-dimensional document embeddings produced in the first stage. UMAP is a well-established non-linear dimensionality reduction technique, particularly effective for handling high-dimensional datasets like document embeddings. It preserves both local and global data structures, thereby enabling improved clustering in the subsequent phase.

UMAP functions by constructing a weighted graph that represents relationships between neighboring data points in the high-dimensional space, using the following formula:

$$u_{ij} = exp\left(-\frac{d(x_i,x_j)-p_i}{\sigma_i}\right) \quad (2)$$

where, d (x_i, x_j) is the distance between data points x_i and x_j, p_i is the distance to the nearest neighbor of x_i, ensuring that each point has at least one strong connection, and σ_i scales the neighborhood size.

Next, UMAP optimizes the low-dimensional layout by minimizing a cross-entropy loss:

$$C = \sum_{i \neq j} u_{ij} \, log\left(\frac{u_{ij}}{v(y_i, y_j)}\right) \quad (3)$$

where, v (y_i, y_j) models the probability of connectivity in the lower-dimensional space.

UMAP's capability to maintain the structural integrity of high-dimensional data while reducing its complexity is critical to the effectiveness of the BERTopic algorithm. By enabling efficient clustering of semantically similar documents in a low-dimensional space, UMAP ensures that the topics identified in the final stage of BERTopic remain coherent and accurately reflect the underlying structure of the document corpus.

3) Hierarchical density-based spatial clustering of applications with noise (HDBSCAN): In the final stage of the BERTopic algorithm, document embeddings, reduced in dimensionality using UMAP, are clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). As an enhancement of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, HDBSCAN offers the ability to detect clusters with varying densities while effectively managing noise. These capabilities are particularly important for handling text data, which often exhibits complex distributions.HDBSCAN is a density-based clustering algorithm capable of adapting to varying densities, enabling the detection of clusters without needing to predefine their number. It determines the core distance for each data point, which is calculated as the distance to its k-th nearest neighbor:

$$core\_dist = d(x_i, x_{k-nearest}) \quad (4)$$

It then defines the mutual reachability distance between point's x_i and x_j:

$$mutual\_reach\_dist(x_i, x_j) = max\left(core\_dist(x_i), core\_dist(x_j), d(x_i, x_j)\right) \quad (5)$$

HDBSCAN constructs a minimum spanning tree and extracts the most stable clusters based on persistence, classifying outliers to reduce noise. This flexibility in handling varied densities and outliers makes it ideal for detecting distinct topics in BERTopic [34].

HDBSCAN is integral to the final stage of BERTopic, where it clusters document embeddings into coherent groups representing distinct topics. Its hierarchical and density-based methodology, coupled with the ability to manage noise and variable-density clusters, makes it particularly suited to the complexities of textual data. By combining UMAP's dimensionality reduction with HDBSCAN's clustering functionality, BERTopic effectively extracts high-quality topics from extensive document datasets.

E. Machine Learning

This study employed six well-established machine learning classification algorithms to evaluate the impact of incorporating textual data, such as clinical notes and pathology reports, in predicting ICU patient mortality. These algorithms were selected based on their demonstrated effectiveness in handling heterogeneous data and addressing complex predictive tasks. The subsequent sections present a detailed overview of the methodologies utilized, including a brief description of each algorithm along with its respective parameter configurations. This comparative analysis aims to highlight the relative strengths of different models in leveraging both structured and unstructured clinical data for enhanced predictive performance. Table II provides a comprehensive summary of the machine learning algorithms employed in this study, detailing their theoretical foundations, optimized parameter settings, including ensemble sizes and kernel types, while maintaining default values for all other parameters to ensure reproducibility.

F. Evaluation Criteria and Metrics

To comprehensively assess the impact of incorporating both structured and semi-structured data on ICU patient mortality prediction, this study employed five key evaluation metrics: AUROC, specificity, sensitivity, precision, and F1-score. These metrics were chosen to ensure a well-rounded assessment of model performance. Table III illustrates the confusion matrix, which forms the basis for calculating these evaluation measures.

Each metric offers distinct insights into model effectiveness. AUROC serves as an indicator of overall classification performance, providing a measure of the model's ability to distinguish between classes. Precision evaluates the correctness of positive predictions, ensuring that identified positive cases are accurate. Sensitivity (recall) gauges the model's capability to correctly detect true positive cases, while specificity determines its effectiveness in identifying true negatives. The F1-score, a harmonic mean of precision and recall, ensures a balanced evaluation by considering both false positives and false negatives. Together, these metrics provide a rigorous framework for comparing models that integrate numerical and text-based clinical data in ICU patient outcome prediction.

TABLE II. SELECTED ALGORITHM AND PARAMETER SETTING

Algorithm	Description	Parameters
Adaboost	A boosting-based ensemble learning algorithm that enhances classification performance by sequentially combining multiple weak learners. It dynamically adjusts the weights of misclassified instances to improve overall model accuracy and robustness [35, 36].	n_estimators=50, learning_rate=1, base_estimator=DecisionTreeClassifier(max_depth=1)
Gradient Boosting (GB)	A powerful ensemble method that constructs models in a sequential manner, where each new model is trained to correct the errors of its predecessors. It leverages gradient descent optimization to minimize the loss function and refine predictive accuracy [37, 38].	n_estimators=100, learning_rate=1.0, max_depth=1, loss='deviance'
Support Vector Classification (SVC)	A classification technique designed to identify the optimal decision boundary that maximizes class separation in high-dimensional feature spaces. This approach is particularly effective in capturing complex, non-linear relationships between variables [39, 40].	kernel='rbf', C=1.0, gamma='auto'
Bagging	A variance-reducing ensemble method that enhances model stability by training multiple classifiers on randomly sampled subsets of the dataset. The final prediction is obtained by aggregating individual model outputs, leading to improved generalization [41, 42].	base_estimator=DecisionTreeClassifier(), n_estimators=500, max_samples=100
Logistic Regression (LR)	A widely applied statistical model for binary classification tasks. It efficiently estimates the probability of class membership and is frequently utilized for risk assessment due to its interpretability and computational efficiency [43, 44].	solver='sag', penalty='l2', C=1.0
Multi-Layer Perceptron Classifier (MLPClassifier)	A feedforward artificial neural network composed of multiple interconnected layers. It employs backpropagation to adjust weights iteratively, allowing it to capture complex, non-linear patterns within data [45, 46].	hidden_layer_sizes=(13,13,13), max_iter=1000, activation='relu', solver='adam'

TABLE III. CONFUSION MATRIX

		Prediction	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Precision = PPV = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = TPR = \frac{TP}{TP+FN} \tag{7}$$

$$F-score = \frac{2*Precision*Recall}{Precision+Recall} \tag{8}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{9}$$

1) Precision: This metric quantifies the proportion of correctly classified positive instances out of all instances predicted as positive. It provides an indication of how reliable the model is in making positive predictions.

2) Recall: Represents the fraction of actual positive cases that the model successfully identifies. It serves as a measure of the model's effectiveness in capturing true positives within the dataset.

3) F1-Score: Defined as the harmonic mean of precision and recall, the F1-score balances the trade-off between these two metrics, ensuring a comprehensive evaluation of classification performance.

4) Accuracy: Expresses the proportion of correctly classified samples, encompassing both positive and negative cases, relative to the entire dataset. It provides an overall assessment of the model's classification performance.

5) AUROC (Area under the receiver operating characteristic curve): A metric that assesses the model's discriminatory power across varying classification thresholds. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), where TPR measures the model's ability to detect actual positive cases, and FPR quantifies the proportion of false positives. AUROC values range from 0 to 1, with higher scores indicating stronger overall classification capability.

III. RESULTS

A. Analysis of Textual Data

The NOTEEVENTS table in the MIMIC-III database contains a vast collection of textual records documenting ICU patient care throughout their hospital stays. In this study, which focuses on mortality prediction, the NOTEEVENTS table was the primary source of text-based information. By analyzing discharge summaries, word clouds were generated to identify key patient characteristics, providing clinical insights that contribute to more effective condition monitoring and management. To extract meaningful features from the clinical text, BERTopic was employed to generate key topics, which were subsequently incorporated as input variables in the predictive models. The process of determining the optimal number of topics and refining the predictive model was guided by methodologies proposed by Baird et al. and Abuzayed et al. [47, 48]. Our findings indicate that ten topics resulted in the highest predictive accuracy. As outlined in Table IV, and visualized in Fig. 2, these ten topics were derived from the processed textual data and linked to specific keywords representing essential patient conditions and medical interventions.

The process of topic generation also allowed for the identification of hidden patterns within the clinical narratives, which may not be immediately apparent through structured data alone. For example, the frequency distribution of clinical descriptors, such as acute symptoms, procedural terms, or medication classes revealed clusters of patient conditions that aligned with known ICU syndromes, thereby validating the clinical relevance of the extracted topics. The identified topics covered various clinical domains, including cardiovascular health, head trauma management, abdominal disorders, and medication prescriptions. The x-axis in each subplot represents the magnitude of the class-based TF-IDF (c-TF-IDF) score, indicating the importance of each word within its respective topic. By leveraging these extracted topics, the model achieved improved predictive performance, offering valuable support for clinical decision-making and patient outcome forecasting.

TABLE IV. IDENTIFIED VARIABLE TOPICS AND ASSOCIATED KEYWORDS

No.	Topic	Keywords
0	Cardiovascular Health	chest, aortic, cardiac, artery, ventricular, capsule, heart, coronary, glucose, mitral, allergies, pulmonary, edema, acute, systolic, aspirin, hypertension, wall
1	Head Trauma and Treatment	head, hemorrhage, fracture, capsule, acute, allergies, seizure, bid, impression, stroke, artery, frontal, mm, midline, surgical, glucose
2	Abdominal and Pancreatic Disorders	abdominal, pancreatitis, biliary, fluid, bile, bowel, stent, acute, abdomen, pancreatic, glucose, surgery, diet, liver, surgical, cholangitis, gallbladder, urine, tube
3	Renal and Urinary System Health	renal, urology, urine, bladder, kidney, stone, tube, prostate, capsule, glucose, hydronephrosis, cancer, pod, foley, surgical, catheter, allergies, fluid
4	Esophageal and Chest Medical Topics	tube, esophageal, esophagus, pod, chest, feeding, cancer, removed, feeds, site, gastric, swallow, incision, drain, glucose, invasive, diet, liquid, surgical
5	Medication Prescriptions and Medical Guidance	bid, capsule, identifier, ec, qd, allergies, disposition, completed, facility, aspirin, tid, attending first, extended, dictated, sodium, medquist, solution, docusate
6	Breast Cancer and Related Surgeries	breast, flap, cancer, carcinoma, pod, squamous, tube, neck, site, cell, plastic, postoperative, metastatic, floor, surgery, surgical, bilateral, mouth, drain
7	Carotid Artery Narrowing and Medical Interventions	carotid, stenosis, artery, ica, stent, stenting, internal, Plavix, qd, pressure, aspirin, angiography, intact, cardiac, cad, chest, stroke, bruit, bilaterally, surgery
8	Hand and Finger-Related Issues	finger, hand, repair, ring, radial, injury, middle, capsule, plastic, forearm, wrist, distal, long, clinic, surgery, joint, postoperatively, saw, sensation, signs
9	Gastric and Obesity-Related Issues	gastric, diet, obesity, weight, bypass, surgery, medication, surgical, drainage, postoperative, advanced, apnea, severe, abdominal, sleep, incisions, leak, ten

Additionally, the topic-word distributions help demonstrate how certain clinical concepts frequently co-occur, suggesting possible interactions between patient comorbidities and acute ICU conditions. Such co-occurrence patterns play a crucial role in mortality prediction, as they offer a richer representation of patient complexity compared to isolated numerical features. The application of BERTopic successfully extracted ten key topics from ICU patient data, providing valuable insights into the diverse and complex clinical conditions observed in intensive care settings. Each identified topic was associated with a specific set of keywords, enabling the seamless integration of both structured and semi-structured data into predictive modeling. This approach underscored the efficacy of combining NLP with machine learning to enhance clinical decision-making and optimize patient care strategies.

Moreover, BERTopic's ability to represent patients through probabilistic topic distributions makes it particularly well-suited for capturing subtle variations in disease severity and clinical presentation. This probabilistic representation allows the model to quantify the degree to which a patient aligns with multiple clinical conditions simultaneously—a feature that is especially important in the ICU, where patients often present with multiple overlapping diagnoses. By transforming semi-structured textual data into meaningful topics and encoding them as probabilistic representations, this study effectively leveraged the full spectrum of information available within electronic health records (EHRs). This methodology demonstrated a significant advancement over traditional predictive models, which typically depend solely on structured numerical data, thereby emphasizing the potential of text-driven analytics in critical care research. Taken together, these findings highlight that narrative clinical documentation is not merely supplementary information but a valuable and often underutilized component of ICU data. Integrating NLP-derived insights into predictive modeling enables a more holistic understanding of patient status, supporting the development of more sensitive and context-aware risk prediction tools.

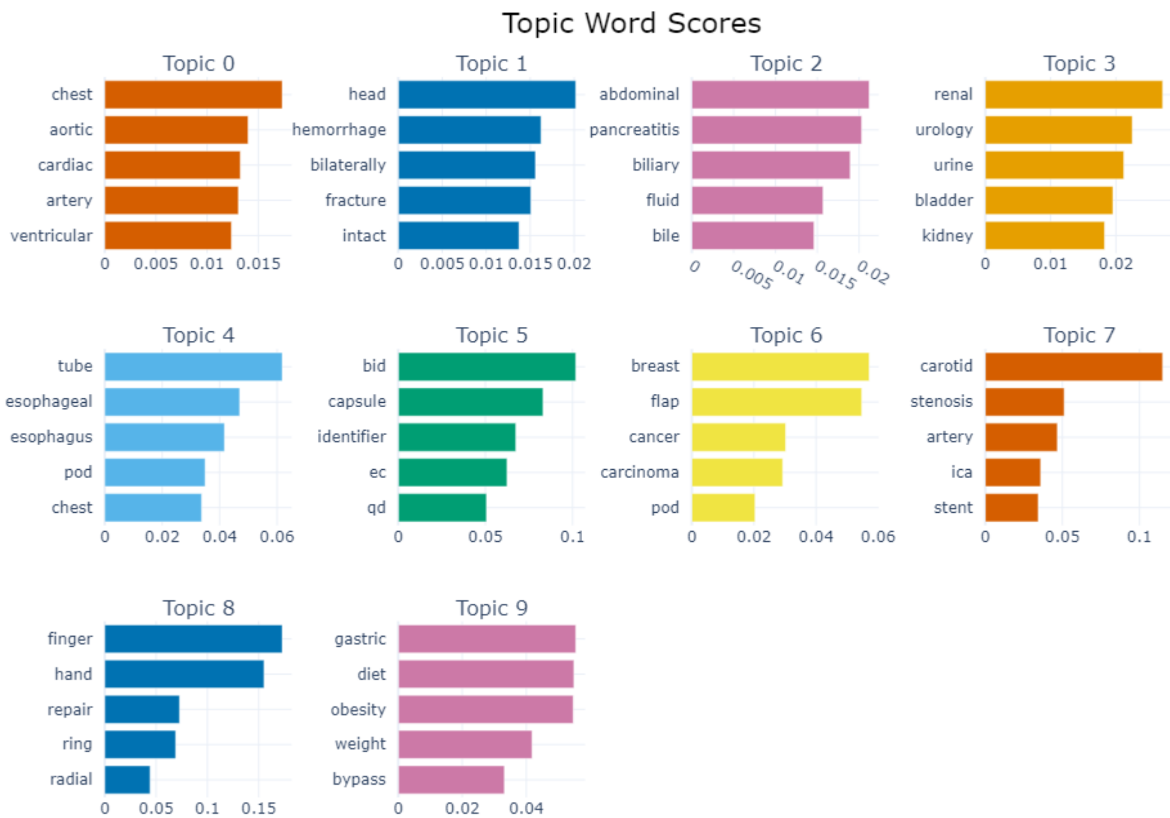

Fig. 2. Topic word scores across identified medical themes in ICU patient data.

B. *Prediction of the Mortality*

To evaluate model performance, this study employed k-fold cross-validation, a widely adopted technique that partitions the dataset into k equal subsets and iteratively trains and validates the model across all folds. By averaging the results across multiple iterations, this method mitigates bias and enhances predictive reliability. However, despite its advantages in reducing overfitting and improving generalization, k-fold cross-validation is computationally demanding [49, 50]. In this study, a 10-fold cross-validation approach was implemented to develop mortality prediction models for 2 days, 3 days, 1 month, and 1 year post-ICU admission, utilizing structured and semi-structured data collected within the first 24 hours of patient admission.

This design allows the model to repeatedly learn from different subsets of patients, improving its ability to generalize across heterogeneous ICU populations where disease severity and documentation patterns vary widely. By incorporating both short-term and long-term mortality outcomes, the study provides a comprehensive assessment of how early clinical information contributes to risk stratification across different temporal horizons.

To build the predictive models, we combined features derived from both structured and unstructured data. The 16 selected quantitative variables, capturing physiological measurements and demographic information, formed the structured data component. For the unstructured data component, the BERTopic model generated 10 distinct topics from the clinical notes (as detailed in Table IV). Each patient's textual data was then represented by these topics, potentially using the probability distribution across the 10 topics or a categorical variable indicating the dominant topic. These topic-derived features were then concatenated with the 16 quantitative features to create a comprehensive feature vector for each patient. This combined feature set was used as input for the six machine learning classifiers described. This integration strategy enables the model to capture complementary sources of information: structured data reflect measurable physiological conditions, while textual data provide implicit clinical reasoning and contextual nuances recorded by healthcare providers. Notably, topic distributions generated by BERTopic often highlight symptom clusters, impressions of clinical instability, or early differential diagnoses, all of which offer predictive signals not present in numeric variables.

Crucially, to enable this comparative analysis evaluating the impact of text features, the patient cohort included in the modeling was necessarily limited to those individuals possessing available and processable clinical notes within the NOTEEVENTS dataset. Comorbidities are implicitly captured within both the quantitative variables (e.g., lab results reflecting organ function) and the clinical notes from which the topics were derived. This also implies that the inclusion of text-based features may reduce noise introduced by incomplete structured data, as clinical notes frequently summarize underlying chronic conditions that may not be fully captured in early laboratory measurements.

For model training and evaluation, the dataset was divided into 80% training data and 20% testing data to ensure robust performance assessment. A series of statistical analyses were conducted to examine the impact of integrating structured EHR data with textual features on ICU mortality prediction. The evaluation framework incorporated five key performance metrics: AUROC, accuracy, precision, recall, and F1-score, providing a comprehensive assessment of predictive capability and model robustness. Table V presents a summary of the model's predictive performance across different time intervals, demonstrating its effectiveness in capturing patient mortality risk over various post-admission periods. Using multiple performance metrics is essential because ICU mortality prediction involves imbalanced outcomes where AUROC alone may not fully reflect the model's clinical utility. Precision and recall help reflect how well the model identifies high-risk individuals while avoiding false alarms, which is crucial for ICU resource allocation. The consistent improvements across several metrics further indicate that textual features enhance not only discrimination but also the stability and reliability of the predictive models across different classifiers.

TABLE V. PERFORMANCE OF THE MODELS

		AdaBoost	**Gradient Boosting**	**SVC**	**Bagging**	**Logistic Regression**	**MLP Classifier**
Without BERTopic Data	2-Days	0.8852 ± 0.0041	0.8823 ± 0.0048	0.8997 ± 0.0112	0.8851 ± 0.0142	0.8955 ± 0.0152	0.8457 ± 0.0304
	3-Days	0.8153 ± 0.0157	0.8129 ± 0.0214	0.8252 ± 0.0153	0.8113 ± 0.0179	0.8228 ± 0.0135	0.7998 ± 0.0063
	1-Month	0.7834 ± 0.0038	0.785 ± 0.0057	0.7715 ± 0.0095	0.782 ± 0.0088	0.769 ± 0.0081	0.7801 ± 0.007
	1-Year	0.7862 ± 0.0039	0.7859 ± 0.0025	0.7711 ± 0.0053	0.7832 ± 0.0065	0.7667 ± 0.0033	0.7768 ± 0.007
With BERTopic Data	2-Days	0.901 ± 0.0167	0.8995 ± 0.013	0.9137 ± 0.0054	0.9038 ± 0.0113	0.9097 ± 0.0049	0.8487 ± 0.0199
	3-Days	0.8332 ± 0.0097	0.8314 ± 0.0057	0.8422 ± 0.0089	0.8372 ± 0.0065	0.846 ± 0.0076	0.8063 ± 0.0136
	1-Month	0.7927 ± 0.0003	0.7935 ± 0.0022	0.7705 ± 0.0032	0.7822 ± 0.0071	0.7682 ± 0.0002	0.7957 ± 0.0047
	1-Year	0.7874 ± 0.005	0.7909 ± 0.003	0.7697 ± 0.0082	0.7799 ± 0.0074	0.7652 ± 0.0069	0.7845 ± 0.0012

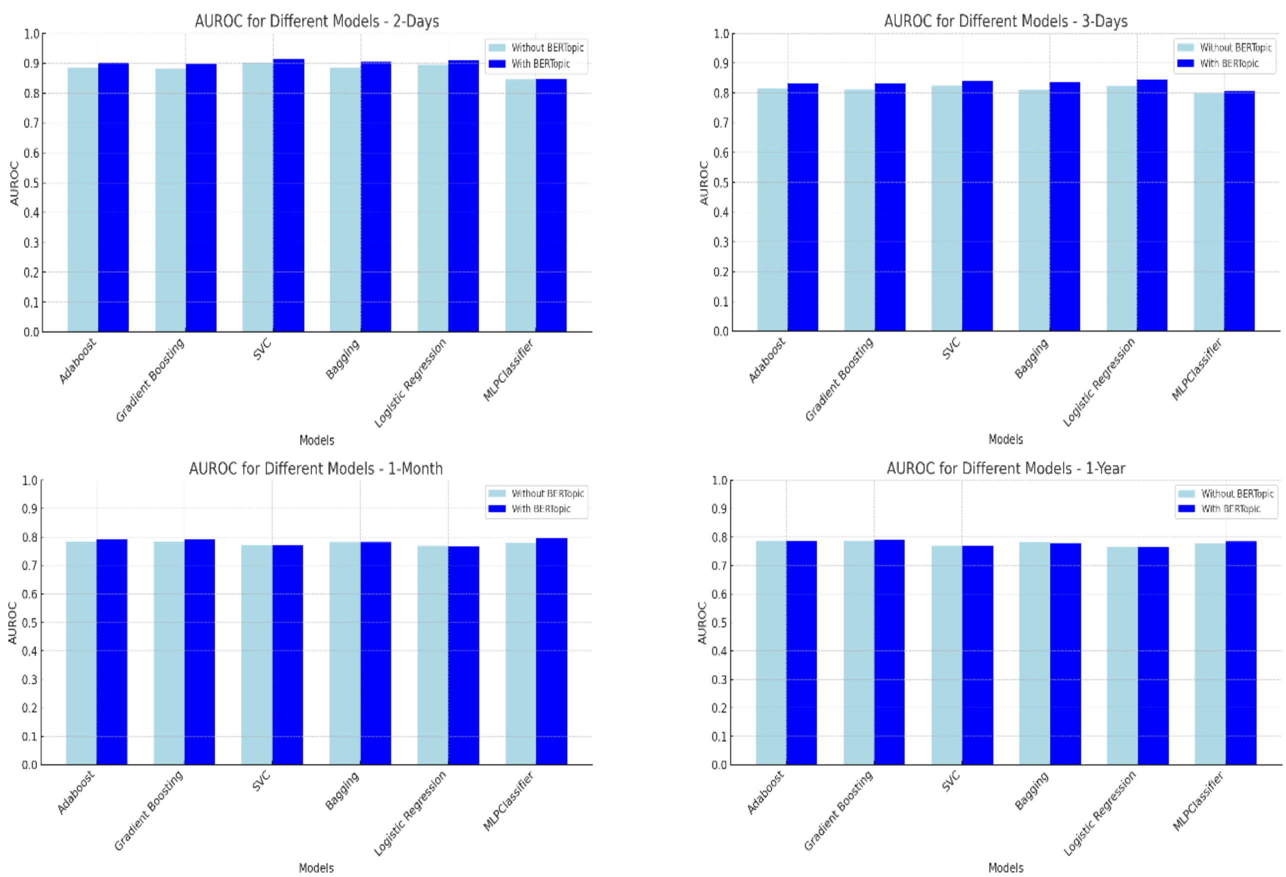

Fig. 3. Bar chart of machine learning model performance: With and without BERTopic data.

The study results demonstrate that AUROC is a vital metric for assessing a model's ability to differentiate between varying risk levels, particularly for identifying high-risk ICU patients who require timely intervention. Incorporating BERTopic data significantly improved the AUROC for all models in short-term predictions (2-day and 3-day). Notably, the AUROC of the SVC model increased from 0.8997 to 0.9137, highlighting the enhanced discriminatory power provided by BERTopic data for identifying high-risk patients. As depicted in Fig. 3, a comparison of machine learning model performance with and without BERTopic data underscores these improvements.

These findings suggest that early clinical notes capture subtle indicators of deterioration, such as mentions of respiratory distress, mental status fluctuation, or clinician concern that structured variables may not immediately reflect. The added predictive value in short-term outcomes highlights the importance of leveraging semi-structured data during the earliest stages of ICU care, when timely risk detection can meaningfully influence clinical decision-making and patient outcomes.

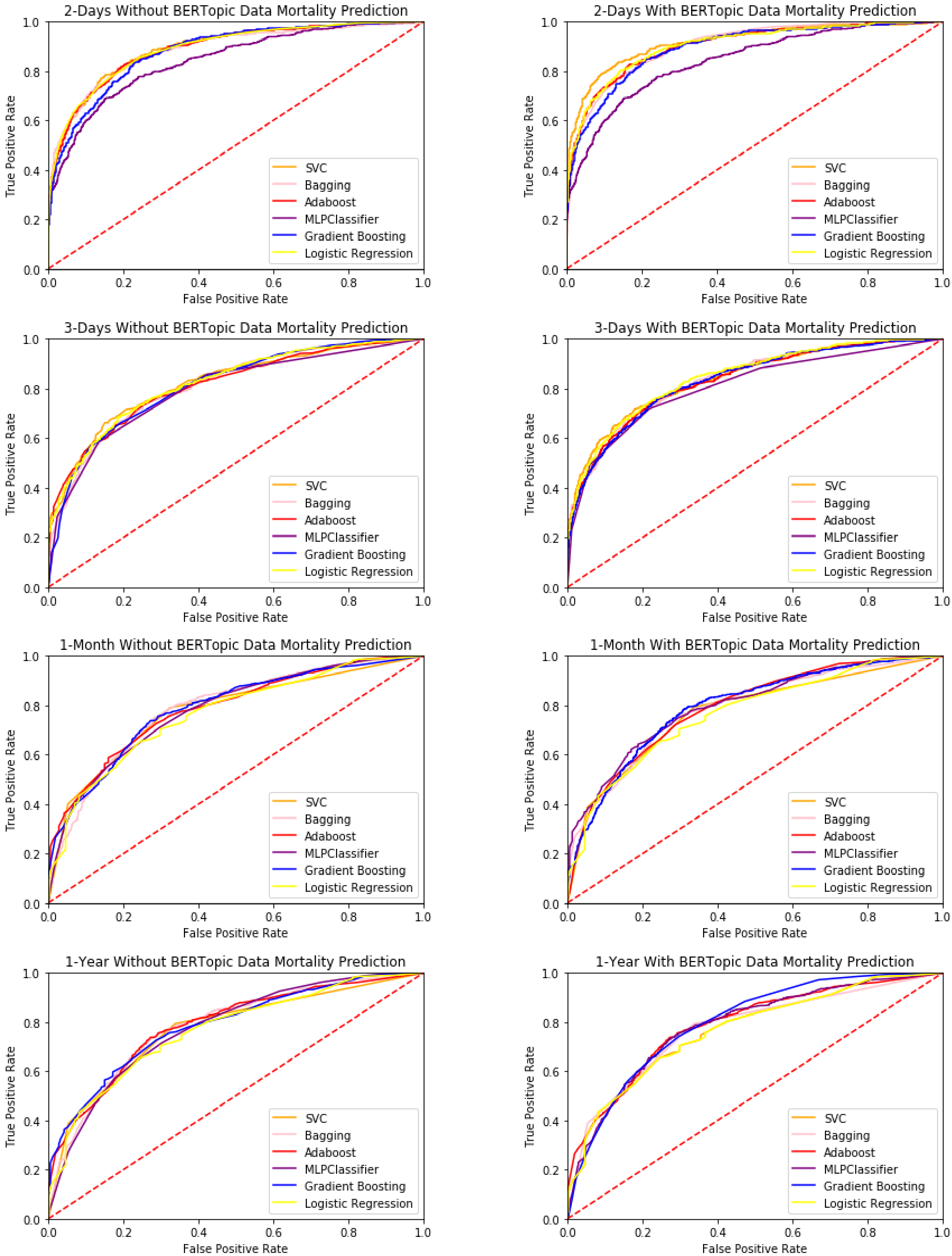

Fig. 4. AUROC curves for different experimental conditions.

The overall AUROC for short-term predictions was higher than that for long-term predictions, suggesting that the models are more effective in distinguishing between high- and low-risk patients over shorter timeframes. However, for long-term predictions (1 month and 1 year), the impact of BERTopic data was less pronounced, with only minor improvements observed in model performance. For instance, the AUROC of the Gradient Boosting model increased slightly from 0.785 to 0.7935, indicating a modest enhancement. Fig. 4 illustrates AUROC curves under different experimental conditions, further corroborating these findings.

These results emphasize the potential of integrating numerical and text-based data to significantly enhance short-term prediction accuracy, particularly for early risk assessment in ICU patients. This insight is invaluable for healthcare professionals in optimizing resource allocation and intervention strategies. However, the limited improvement in long-term predictions indicates challenges in differentiating risk levels over extended periods. Future research should focus on incorporating additional data sources or exploring novel techniques to enhance long-term prediction performance.

In this study, alongside AUROC, which measures the model's ability to discriminate between classes, we evaluated its practical applicability using four additional performance metrics: Accuracy, Recall, Precision, and F1-score. Collectively, these metrics deliver a comprehensive assessment of the models' predictive performance, encompassing both overall accuracy and their effectiveness in precisely identifying high-risk patients. As shown in Table VI, the incorporation of BERTopic data led to significant improvements in model performance, particularly in short-term predictions (2-day and 3-day). For example, the Accuracy of the Gradient Boosting model for 2-day predictions increased from 0.9339 to 0.9409, highlighting the positive impact of integrating text data. Across all models, Accuracy was generally higher for short-term predictions compared to long-term ones, reflecting the models' greater precision in shorter timeframes.

TABLE VI. COMPARISON OF MACHINE LEARNING MODEL PERFORMANCE: WITHOUT AND WITH BERTOPIC DATA

	Time Period	**Metrics**	**AdaBoost**	**Gradient Boosting**	**SVC**	**Bagging**	**Logistic Regression**	**MLP Classifier**
Without BERTopic Data	2-Days	Precision	0.1155 ± 0.0065	0.1257 ± 0.0055	0.0875 ± 0.0057	0.081 ± 0.0056	0.0865 ± 0.0052	0.0956 ± 0.0193
		Recall	0.8433 ± 0.0099	0.8295 ± 0.0099	0.9058 ± 0.0238	0.883 ± 0.0305	0.8977 ± 0.0339	0.7773 ± 0.0722
		F1	0.2032 ± 0.0102	0.2183 ± 0.0081	0.1596 ± 0.0099	0.1484 ± 0.0098	0.1578 ± 0.0092	0.1694 ± 0.0301
		Accuracy	0.9262 ± 0.0045	0.9339 ± 0.0009	0.8937 ± 0.0012	0.8871 ± 0.0019	0.8933 ± 0.0032	0.9126 ± 0.018
	3-Days	Precision	0.1354 ± 0.0105	0.1298 ± 0.0082	0.1092 ± 0.0045	0.1182 ± 0.0092	0.1011 ± 0.0031	0.1122 ± 0.0104
		Recall	0.7335 ± 0.0419	0.7335 ± 0.0497	0.7915 ± 0.0387	0.7447 ± 0.0465	0.8011 ± 0.0317	0.7278 ± 0.0337
		F1	0.2281 ± 0.0133	0.2202 ± 0.0112	0.1918 ± 0.0062	0.2037 ± 0.013	0.1795 ± 0.0047	0.194 ± 0.0144
		Accuracy	0.8935 ± 0.0099	0.8889 ± 0.0072	0.8574 ± 0.0073	0.8751 ± 0.0112	0.8436 ± 0.004	0.8687 ± 0.0225
	1-Month	Precision	0.2729 ± 0.0031	0.2689 ± 0.0068	0.247 ± 0.0023	0.2583 ± 0.01	0.2402 ± 0.0032	0.233 ± 0.0067
		Recall	0.7441 ± 0.0087	0.7532 ± 0.0083	0.7466 ± 0.0252	0.7592 ± 0.016	0.7501 ± 0.0222	0.7941 ± 0.0133
		F1	0.3994 ± 0.0045	0.3963 ± 0.0086	0.3712 ± 0.0055	0.3853 ± 0.0126	0.3639 ± 0.0056	0.3602 ± 0.009
		Accuracy	0.8163 ± 0.0031	0.8115 ± 0.0037	0.7925 ± 0.0038	0.8011 ± 0.0052	0.7847 ± 0.0038	0.7684 ± 0.0035
	1-Year	Precision	0.2778 ± 0.0108	0.2761 ± 0.0121	0.2554 ± 0.012	0.2618 ± 0.0135	0.2446 ± 0.0133	0.247 ± 0.0415
		Recall	0.7579 ± 0.0122	0.7595 ± 0.0141	0.7474 ± 0.0158	0.7711 ± 0.0145	0.7524 ± 0.0076	0.789 ± 0.0571
		F1	0.4063 ± 0.01	0.4047 ± 0.0113	0.3805 ± 0.0133	0.3908 ± 0.0161	0.369 ± 0.0155	0.3727 ± 0.0411
		Accuracy	0.8096 ± 0.0041	0.8078 ± 0.0083	0.7908 ± 0.0038	0.7932 ± 0.0023	0.7786 ± 0.0049	0.7666 ± 0.0467
With BERTopic Data	2-Days	Precision	0.1199 ± 0.0046	0.1399 ± 0.0046	0.085 ± 0.0043	0.0833 ± 0.0062	0.0867 ± 0.0037	0.1007 ± 0.0149
		Recall	0.8727 ± 0.0323	0.8571 ± 0.0268	0.9391 ± 0.0144	0.9196 ± 0.0284	0.9272 ± 0.0144	0.7752 ± 0.0475
		F1	0.2107 ± 0.0072	0.2404 ± 0.0061	0.1558 ± 0.0072	0.1527 ± 0.0105	0.1586 ± 0.0064	0.1776 ± 0.0228
		Accuracy	0.9287 ± 0.0014	0.9409 ± 0.0017	0.8888 ± 0.0036	0.8883 ± 0.0061	0.8925 ± 0.0044	0.9206 ± 0.0102
	3-Days	Precision	0.1426 ± 0.0068	0.1439 ± 0.0042	0.1127 ± 0.0044	0.1238 ± 0.0069	0.1078 ± 0.0047	0.114 ± 0.015
		Recall	0.7716 ± 0.0185	0.7662 ± 0.0108	0.833 ± 0.0198	0.8034 ± 0.0109	0.8518 ± 0.0158	0.7479 ± 0.0555
		F1	0.2407 ± 0.0106	0.2423 ± 0.0064	0.1986 ± 0.0074	0.2144 ± 0.0107	0.1914 ± 0.0077	0.1969 ± 0.0202
		Accuracy	0.8921 ± 0.0015	0.8938 ± 0.0035	0.8511 ± 0.0016	0.8694 ± 0.0039	0.8405 ± 0.0014	0.862 ± 0.0267
	1-Month	Precision	0.2799 ± 0.0088	0.2789 ± 0.0098	0.2488 ± 0.0023	0.2568 ± 0.0005	0.2384 ± 0.0044	0.2894 ± 0.0079
		Recall	0.7613 ± 0.0038	0.7646 ± 0.0103	0.7423 ± 0.0032	0.7626 ± 0.0109	0.7522 ± 0.0039	0.7595 ± 0.0169
		F1	0.4092 ± 0.0092	0.4086 ± 0.0092	0.3726 ± 0.0021	0.3842 ± 0.0011	0.362 ± 0.0053	0.4189 ± 0.0066
		Accuracy	0.8188 ± 0.0035	0.8176 ± 0.005	0.794 ± 0.0052	0.7985 ± 0.0042	0.7816 ± 0.0034	0.826 ± 0.0127
	1-Year	Precision	0.2886 ± 0.0005	0.2891 ± 0.0027	0.2603 ± 0.0077	0.2683 ± 0.0084	0.2475 ± 0.0061	0.2781 ± 0.0051
		Recall	0.7536 ± 0.0155	0.7622 ± 0.0084	0.7427 ± 0.0192	0.7594 ± 0.01	0.7498 ± 0.0161	0.7585 ± 0.0076
		F1	0.4173 ± 0.0019	0.4192 ± 0.0032	0.3855 ± 0.0102	0.3964 ± 0.0105	0.3722 ± 0.0086	0.407 ± 0.0046
		Accuracy	0.8152 ± 0.0049	0.8145 ± 0.0052	0.7921 ± 0.0042	0.7969 ± 0.0059	0.7779 ± 0.003	0.8059 ± 0.0043

Based on the data in Table VI, the Recall metric highlights the models' ability to correctly identify actual positive cases, such as ICU patient mortality. In healthcare, particularly in critical care settings, Recall is crucial as it ensures the detection of all high-risk patients, allowing for timely interventions. The table shows that adding BERTopic data substantially increased Recall in short-term predictions. For example, the SVC model's Recall improved from 0.9058 to 0.9391 for 2-day predictions, while Bagging's Recall rose from 0.883 to 0.9196. However, the improvements were less pronounced for longer timeframes, with relatively stable performance for 1-month and 1-year predictions. This suggests that while BERTopic data significantly enhances short-term predictive performance, its impact on long-term predictions is more limited.

C. Analysis of Variable Importance

A key advantage of the Gradient Boosting (GB) method is its capability to assess the relative contribution of input variables in predictive modeling. This is achieved through the computation of variable importance scores, which rank features based on their influence on model performance. These scores indicate how significantly each variable impacts decision-making within the model, with higher importance assigned to variables frequently utilized in decision tree splits.

In Gradient Boosting, the importance of a variable is determined by evaluating the increase in the decision tree's value at a given split point, adjusted by the number of samples at that node. Several established metrics, including the Gini index, cross-entropy, and information gain, can be used to measure decision tree value changes. This study employed the Gini index to quantify variable significance, ensuring a consistent and interpretable ranking approach. For a comprehensive discussion on variable importance calculations in GB models, readers may refer to Hastie et al. [51]. Table VII presents the importance scores derived from the optimal Gradient Boosting model, which was trained using ICU patient data collected within the first 24 hours post-admission.

Table VII highlights the key variables selected for predicting both short-term and long-term mortality in ICU patients using the GB method. The analysis reveals that different variables contribute uniquely to mortality prediction across various timeframes (2 days, 3 days, 1 month, and 1 year). For short-term mortality (2 and 3 days), the Glasgow Coma Scale (X10) consistently emerges as the most significant variable, underscoring its critical role in assessing patient consciousness and predicting immediate outcomes in ICU settings. Other important indicators for short-term predictions include White Blood Cell Count (X6) and Heart Rate (X2), which are essential for evaluating acute physiological conditions, including infections and potential cardiac issues. In contrast, long-term mortality predictions (1 month and 1 year) are dominated by heart-related factors such as Heart Rate (X2) and Systolic Blood Pressure (X3), reflecting the sustained impact of cardiovascular health on patient outcomes over extended periods. Additionally, age (X1) and temperature (X4) gain prominence in long-term predictions, as chronic health conditions and the regulation of body temperature become increasingly relevant to survival.

TABLE VII. TEN VARIABLES IDENTIFIED FROM 24-HOUR DATASETS

	Order of Variable Importance	Short-Term Mortality		Long-Term Mortality	
		2 Days	*3 Days*	*1 month*	*1 year*
Without BERTopic Data	1	X10	X10	X2	X2
	2	X6	X6	X6	X6
	3	X2	X2	X10	X4
	4	X4	X1	X5	X9
	5	X1	X3	X3	X10
	6	X5	X9	X4	X1
	7	X3	X4	X7	X5
	8	X8	X8	X1	X3
	9	X7	X5	X9	X7
	10	X9	X7	X8	X8
With BERTopic Data	1	X10	X10	X10	X10
	2	X6	X1	X2	X4
	3	X1	X2	X4	X6
	4	X4	X4	X1	X2
	5	X3	X6	X6	X1
	6	X2	TOPIC 1	X5	X9
	7	TOPIC 1	TOPIC 2	X3	X5
	8	TOPIC 0	X9	TOPIC 1	X3
	9	X8	X7	TOPIC 0	TOPIC 1
	10	X9	X3	X8	X8

The integration of topic modeling data (such as Topic 1 and Topic 0) into the variable importance rankings underscores the importance of textual clinical notes in enhancing model performance, particularly for predicting long-term mortality. These topics encompass critical health domains such as cardiovascular health (Topic 0), head trauma (Topic 1), and abdominal and pancreatic disorders (Topic 2). The emergence of these text-derived topics as important predictors, particularly demonstrating notable influence in long-term mortality predictions (as seen for Topic 0 and Topic 1 in Table VII), underscores their value in capturing complex patient narratives and underlying conditions that evolve beyond immediate physiological measurements. The GB method's variable importance analysis underscores the relevance of immediate physiological indicators like GCS and WBC in short-term predictions, while long-term mortality is influenced by a combination of age, cardiovascular health, and text-based clinical information. This analysis provides valuable insights into ICU patient care, allowing healthcare professionals to focus on the most critical variables for improving predictive accuracy across different time horizons.

IV. DISCUSSION

A. Principal Findings

This study illustrates the transformative potential of combining text mining techniques, such as BERTopic, with machine learning models to predict ICU patient outcomes. The key findings of the research are as follows.

1) Enhanced short-term predictions with BERTopic integration: Incorporating BERTopic data into machine learning models significantly improved short-term mortality predictions (2-day and 3-day). For example, the AUROC of the SVC model increased from 0.8997 to 0.9137, indicating an enhanced ability to discriminate between high- and low-risk patients. This enhancement stems from BERTopic's ability to capture rich contextual information embedded within clinical notes, such as clinician assessments of severity, nuanced descriptions of patient status, and evolving treatment plans, which often provide complementary insights beyond structured physiological data. This is particularly crucial for short-term outcomes where acute changes documented in free text are highly indicative of immediate risk.

2) Superior predictive performance in short-term predictions compared to long-term: Across all models, short-term predictions consistently outperformed long-term predictions (1-month and 1-year) in metrics such as AUROC, accuracy, recall, and F1-score. The reduced predictive accuracy over longer timeframes suggests that while models effectively identify immediate risks, long-term mortality predictions require additional variables or more refined methodologies to enhance performance [52].

3) Significance of physiological variables in predictive models: Analysis of variable importance using the Gradient Boosting method revealed that physiological variables, such as the Glasgow Coma Scale, White Blood Cell Count, Heart Rate, and Temperature, are essential predictors of both short-term and long-term ICU patient mortality [53]. The critical role of these variables underscores the necessity of closely monitoring vital signs in critical care environments.

4) Improved mortality prediction through text data integration: Incorporating topic modeling data derived from clinical notes, such as those related to cardiovascular health and head trauma, significantly enhanced mortality predictions [54]. This finding demonstrates the value of leveraging semi-structured and unstructured data through text mining. By converting free text into quantifiable topic features, our approach unlocked clinical insights related to complex patient conditions and care trajectories, often overlooked by traditional models relying solely on numerical data [55].

By integrating text mining and machine learning, this approach provides a deeper understanding of patient conditions, enhancing predictive accuracy and supporting more effective clinical decision-making. Leveraging both structured and unstructured data improves the ability to capture critical insights, facilitating personalized care and timely interventions in the complex environment of critical care settings. These improvements directly contribute to resilient healthcare by strengthening early risk recognition and enabling more adaptive and timely responses in critical care settings.

B. Limitations

Despite the encouraging findings, this study has several limitations that must be acknowledged. While the MIMIC-III database is comprehensive, it originates from a single healthcare system, which may limit the generalizability of the developed models. Recalibration may be necessary when applying these models to different ICU settings, where variations in patient populations, clinical practices, and treatment protocols can influence model performance in real-world applications [56]. Differences in patient demographics and resource availability further highlight the need for external validation. Furthermore, the exclusion criteria based on missing data percentages (>30% for patient records, >40% for predictors) might introduce selection bias. A formal analysis comparing the characteristics of the excluded patient population with the included cohort was not performed in this study and represents a limitation.

Moreover, direct numerical comparison of performance metrics (e.g., AUROC, Accuracy) with previously published studies, including those also utilizing MIMIC-III, is often challenging and potentially misleading. This difficulty arises from inherent variations in specific patient cohort selection criteria (especially regarding text availability), the precise feature sets utilized, differing text processing techniques (e.g., alternative NLP models or embeddings), distinct prediction target definitions, and varying evaluation protocols across research efforts. Additionally, although the incorporation of text data improved model accuracy, this study analyzed only a subset of the textual information available in the MIMIC-III database. Key clinical data, such as imaging reports, physician progress notes, and detailed medication histories, were not included in the analysis. Expanding the scope of semi-structured data could enhance the predictive accuracy of models, particularly for long-term mortality predictions.

While short-term mortality predictions showed substantial improvements, the performance gains for long-term predictions (1-month and 1-year) were comparatively modest. Long-term outcomes are inherently complex and likely require the inclusion of additional variables, such as clinical, social, and environmental factors, to enhance predictive accuracy [57, 58]. Furthermore, long-term trajectories are influenced by factors extending beyond the initial 24-hour ICU data and clinical notes, such as progression of chronic conditions, intercurrent events, post-discharge care, and socioeconomic factors not fully captured in MIMIC-III. The predictive signal from initial text data may also diminish over extended periods. These findings underscore the necessity of investigating additional long-term predictors to improve modeling in critical care contexts.

The integration of advanced text mining techniques, such as BERTopic, with machine learning models introduces significant computational challenges. Extracting topics from clinical notes and incorporating them into predictive models requires substantial processing power, which may hinder the real-time application of these methods in clinical settings. While cross-validation was employed to reduce the risk of overfitting, concerns remain about the models' ability to generalize to new patient populations, particularly due to the high dimensionality of text data. Future research should focus on enhancing generalizability by validating these models on larger and more diverse datasets (e.g., multi-center data like MIMIC-IV or eICU), as well as employing advanced regularization strategies or transfer learning techniques. Overcoming these challenges will be essential for developing predictive models that are both accurate and robust, enabling their effective deployment in critical care environments [59].

V. Conclusion

This study investigated the integration of advanced text mining techniques, specifically the BERTopic model, with machine learning algorithms to enhance the prediction of ICU patient outcomes. By leveraging semi-structured data from clinical notes alongside structured numerical data, the research demonstrated that combining these data types significantly improves the accuracy and robustness of predictive models, especially for short-term mortality prediction. These findings highlight several key conclusions:

The incorporation of BERTopic data significantly enhanced short-term mortality prediction, particularly for 2-day and 3-day outcomes. By integrating semi-structured clinical notes into machine learning models, the study achieved improved AUROC and accuracy scores, underscoring the critical role of text data in early ICU care. Often underutilized in predictive modeling, semi-structured data provided essential clinical context that improved the models' ability to identify high-risk patients, enabling more timely and effective interventions [60]. This contextual information, capturing clinical reasoning and nuances absent in structured data, served as a valuable complement to traditional predictors. Additionally, physiological variables such as the Glasgow Coma Scale, White Blood Cell Count, and Heart Rate were identified as crucial predictors for both short-term and long-term mortality. These findings emphasize the importance of continuous monitoring of these indicators to improve outcomes for ill patients. Although the predictive improvements for long-term mortality (1-month and 1-year) were relatively modest, the integration of text-based clinical insights through topic modeling provided valuable contributions to understanding patient conditions over extended periods. These findings highlight the potential of natural language processing techniques, such as BERTopic, to augment structured data and improve long-term predictive accuracy.

This study also underscores the broader applications of AI and big data in clinical decision-making and resource allocation, offering healthcare providers advanced tools to enhance patient outcomes and optimize ICU resource management [61]. By improving early risk recognition and supporting more adaptive clinical responses, these AI-driven predictive frameworks contribute directly to resilient healthcare in ICU environments. Looking ahead, future research should prioritize larger datasets, cross-institutional validation, and the incorporation of additional clinical variables and potentially longitudinal or richer textual sources (e.g., full progress notes and radiology reports) to further refine and generalize predictive models for ICU patient outcomes [62]. The integration of structured numerical data and semi-structured text, processed using advanced text mining techniques such as BERTopic, represents a promising avenue for improving ICU outcome prediction and strengthening data-driven clinical practice. As multimodal AI continues to evolve, its capacity to strengthen resilience in critical care systems through earlier detection, more precise triage, and timely intervention will become increasingly central to improving patient outcomes.

References

[1] R. M. McAdams, R. Kaur, Y. Sun, H. Bindra, S. J. Cho, and H. Singh, "Predicting clinical outcomes using artificial intelligence and machine learning in neonatal intensive care units: a systematic review," Journal of Perinatology, vol. 42, no. 12, pp. 1561-1575, 2022.

[2] S. Lee et al., "Unlocking the potential of electronic health records for health research," International Journal of Population Data Science, vol. 5, no. 1, 2020.

[3] S. Aminizadeh et al., "Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service," Artificial Intelligence in Medicine, vol. 149, p. 102779, 2024.

[4] A. Olalekan Kehinde, "Leveraging machine learning for predictive models in healthcare to enhance patient outcome management," Int Res J Mod Eng Technol Sci, vol. 7, no. 1, p. 1465, 2025.

[5] O. Abdolazimi, M. Salehi Esfandarani, M. Salehi, D. Shishebori, and M. Shakhsi-Niaei, "Development of sustainable and resilient healthcare and non-cold pharmaceutical distribution supply chain for COVID-19 pandemic: a case study," The International Journal of Logistics Management, vol. 34, no. 2, pp. 363-389, 2023.

[6] I. Li et al., "Neural natural language processing for unstructured data in electronic health records: a review," Computer Science Review, vol. 46, p. 100511, 2022.

[7] J. Roca, A. Tenyi, and I. Cano, "Paradigm changes for diagnosis: using big data for prediction," Clinical Chemistry and Laboratory Medicine (CCLM), vol. 57, no. 3, pp. 317-327, 2019.

[8] C. H. Guo, M. L. Lu, and J. F. Chen, "An evaluation of time series summary statistics as features for clinical prediction tasks," Bmc Medical Informatics and Decision Making, vol. 20, no. 1, Mar 2020, Art no. 48, doi: 10.1186/s12911-020-1063-x.

[9] S. G. Emami, V. Lorenzoni, and G. Turchetti, "Towards resilient healthcare systems: a framework for crisis management," International journal of environmental research and public health, vol. 21, no. 3, p. 286, 2024.

[10] V. M. Chigboh, S. J. C. Zouo, and J. Olamijuwon, "Predictive analytics in emergency healthcare systems: A conceptual framework for reducing response times and improving patient care," World Journal of Advanced Pharmaceutical and Medical Research, vol. 7, no. 2, pp. 119-127, 2024.

[11] F. Goodarzian, P. Ghasemi, A. Gunasekaren, A. A. Taleizadeh, and A. Abraham, "A sustainable-resilience healthcare network for handling COVID-19 pandemic," Annals of operations research, vol. 312, no. 2, pp. 761-825, 2022.

[12] B. J. Marafino, G. J. Escobar, M. T. Baiocchi, V. X. Liu, C. C. Plimier, and A. Schuler, "Evaluation of an intervention targeted with predictive analytics to prevent readmissions in an integrated health system: observational study," bmj, vol. 374, 2021.

[13] S. S. Siddiqui et al., "Evaluation and validation of four scoring systems: the apache IV, saps III, MPM0 II, and ICMM in critically ill cancer patients," Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine, vol. 24, no. 4, p. 263, 2020.

[14] A. Schoe, F. Bakhshi-Raiez, N. de Keizer, J. T. van Dissel, and E. de Jonge, "Mortality prediction by SOFA score in ICU-patients after cardiac surgery; comparison with traditional prognostic–models," BMC anesthesiology, vol. 20, pp. 1-8, 2020.

[15] D. van de Sande, M. E. van Genderen, J. Huiskens, D. Gommers, and J. van Bommel, "Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit," Intensive care medicine, vol. 47, pp. 750-760, 2021.

[16] F. Tetschke, U. Schneider, E. Schleussner, O. W. Witte, and D. Hoyer, "Assessment of fetal maturation age by heart rate variability measures

using random forest methodology," Comput. Biol. Med., vol. 70, pp. 157-162, 2016.

[17] M. Hashir and R. Sawhney, "Towards unstructured mortality prediction with free-text clinical notes," Journal of Biomedical Informatics, vol. 108, Aug 2020, Art no. 103489, doi: 10.1016/j.jbi.2020.103489.

[18] V. Ntinopoulos et al., "Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation," BMJ Health & Care Informatics, vol. 32, no. 1, p. e101139, 2025.

[19] M. S. Tootooni, K. S. Pasupathy, H. A. Heaton, C. M. Clements, and M. Y. Sir, "CCMapper: An adaptive NLP-based free-text chief complaint mapping algorithm," Comput. Biol. Med., vol. 113, p. 103398, 2019.

[20] E. E. Uslu, E. Sezer, and Z. A. Guven, "NLP-Powered Insights: A Comparative Analysis for Multi-Labeling Classification with MIMIC-CXR Dataset," IEEE Access, 2024.

[21] N. Zahra, H. Tanveer, and N. Batool, "Technological advancement in COVID-19 rehabilitation: Therapists' views," Int J Multidiscip Res Growth Eval, 2025.

[22] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural language processing of clinical notes on chronic diseases: systematic review," JMIR Med. Inf., vol. 7, no. 2, p. e12239, 2019.

[23] L. Gan et al., "Experimental comparison of three topic modeling methods with LDA, Top2Vec and BERTopic," in International Symposium on Artificial Intelligence and Robotics, 2023: Springer, pp. 376-391.

[24] R. Raman, D. Pattnaik, L. Hughes, and P. Nedungadi, "Unveiling the dynamics of AI applications: A review of reviews using scientometrics and BERTopic modeling," Journal of Innovation & Knowledge, vol. 9, no. 3, p. 100517, 2024.

[25] M. Borčin and J. M. Jose, "Optimizing bertopic: Analysis and reproducibility study of parameter influences on topic modeling," in European conference on information retrieval, 2024: Springer, pp. 147-160.

[26] N. Khodeir and F. Elghannam, "Efficient topic identification for urgent MOOC Forum posts using BERTopic and traditional topic modeling techniques," Education and Information Technologies, pp. 1-27, 2024.

[27] A. E. Johnson et al., "MIMIC-III, a freely accessible critical care database," Sci Data, vol. 3, p. 160035, May 24 2016, doi: 10.1038/sdata.2016.35.

[28] T. Gangavarapu, A. Jayasimha, G. S. Krishnan, and S. S. Kamath, "Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes," Knowledge-Based Systems, vol. 190, Feb 2020, Art no. 105321, doi: 10.1016/j.knosys.2019.105321.

[29] Y. Y. Tang et al., "Association of Systemic Immune-Inflammation Index With Short-Term Mortality of Congestive Heart Failure: A Retrospective Cohort Study," (in English), Front. Cardiovasc. Med., Article vol. 8, p. 15, Nov 2021, Art no. 753133, doi: 10.3389/fcvm.2021.753133.

[30] S. Davidson et al., "Day-to-day progression of vital-sign circadian rhythms in the intensive care unit," (in English), Crit. Care, Article vol. 25, no. 1, p. 13, Apr 2021, Art no. 156, doi: 10.1186/s13054-021-03574-w.

[31] M. Sayed, D. Riano, and J. Villar, "Predicting Duration of Mechanical Ventilation in Acute Respiratory Distress Syndrome Using Supervised Machine Learning," Journal of Clinical Medicine, vol. 10, no. 17, Sep 2021, Art no. 3824, doi: 10.3390/jcm10173824.

[32] K. Alghatani, N. Ammar, A. Rezgui, and A. Shaban-Nejad, "Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation," JMIR Med. Inf., vol. 9, no. 5, May 2021, Art no. e21347, doi: 10.2196/21347.

[33] J. T. Kent, "Information gain and a general measure of correlation," Biometrika, vol. 70, no. 1, pp. 163-173, 1983.

[34] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.

[35] D. H. Kim, J. Y. Choi, and Y. M. Ro, "Region based stellate features combined with variable selection using AdaBoost learning in mammographic computer-aided detection," Comput. Biol. Med., vol. 63, pp. 238-250, Aug 2015, doi: 10.1016/j.compbiomed.2014.09.006.

[36] Y. W. Lee, J. W. Choi, and E. H. Shin, "Machine learning model for predicting malaria using clinical information," Comput. Biol. Med., vol. 129, Feb 2021, Art no. 104151, doi: 10.1016/j.compbiomed.2020.104151.

[37] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189-1232, Oct 2001, doi: 10.1214/aos/1013203451.

[38] S. B. Keser and K. Keskin, "A gradient boosting-based mortality prediction model for COVID-19 patients," Neural Computing and Applications, vol. 35, no. 33, pp. 23997-24013, 2023.

[39] S. Nanayakkara et al., "Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study," (in English), PLos Med., Article vol. 15, no. 11, p. 16, Nov 2018, Art no. e1002709, doi: 10.1371/journal.pmed.1002709.

[40] G. Akbari et al., "Frailty Level Classification of the Community Elderly Using Microsoft Kinect-Based Skeleton Pose: A Machine Learning Approach," (in English), Sensors, Article vol. 21, no. 12, p. 20, Jun 2021, Art no. 4017, doi: 10.3390/s21124017.

[41] S. Ali, A. Majid, S. G. Javed, and M. Sattar, "Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data," (in English), Comput. Biol. Med., Article vol. 73, pp. 38-46, Jun 2016, doi: 10.1016/j.compbiomed.2016.04.002.

[42] C. K. Sarmah and S. Samarasinghe, "Microarray gene expression: A study of between-platform association of Affymetrix and cDNA arrays," (in English), Comput. Biol. Med., Article vol. 41, no. 10, pp. 980-986, Oct 2011, doi: 10.1016/j.compbiomed.2011.08.007.

[43] D. R. Cox, "The Regression Analysis of Binary Sequences," Journal of the Royal Statistical Society. Series B (Methodological), vol. 20, no. 2, pp. 215-242, 1958.

[44] A. Lavrentieva, E. Kaimakamis, V. Voutsas, and M. Bitzani, "An observational study on factors associated with ICU mortality in Covid-19 patients and critical review of the literature," Scientific Reports, vol. 13, no. 1, p. 7804, 2023.

[45] T. Windeatt, "Accuracy/Diversity and Ensemble MLP Classifier Design," IEEE Transactions on Neural Networks, vol. 17, no. 5, pp. 1194-1211, 2006, doi: 10.1109/TNN.2006.875979.

[46] A. Bokhare, A. Bhagat, and R. Bhalodia, "Multi-layer perceptron for heart failure detection using SMOTE technique," SN Computer Science, vol. 4, no. 2, p. 182, 2023.

[47] Y. W. Lin, Y. Q. Zhou, F. Faghri, M. J. Shawl, and R. H. Campbell, "Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long shortterm memory," Plos One, vol. 14, no. 7, Jul 2019, Art no. e0218942, doi: 10.1371/journal.pone.0218942.

[48] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," in 5th Conference on AI in Computational Linguistics (ACLing), Electr Network, Jun 04-05 2021, vol. 189, in Procedia Computer Science, 2021, pp. 191-194, doi: 10.1016/j.procs.2021.05.096. [Online]. Available: <Go to ISI>://WOS:000684216300023

[49] B. Liu, L. Fang, F. Liu, X. Wang, and K.-C. Chou, "iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach," Journal of Biomolecular Structure and Dynamics, vol. 34, no. 1, pp. 223-235, 2016.

[50] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, and K.-C. Chou, "Identification of real microRNA precursors with a pseudo structure status composition approach," PloS one, vol. 10, no. 3, p. e0121501, 2015.

[51] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY," 2001.

[52] F. Aziz et al., "Short-and long-term mortality prediction after an acute ST-elevation myocardial infarction (STEMI) in Asians: A machine learning approach," PloS one, vol. 16, no. 8, p. e0254894, 2021.

[53] F. C. Bennis et al., "Improving prediction of favourable outcome after 6 months in patients with severe traumatic brain injury using physiological cerebral parameters in a multivariable logistic regression model," Neurocritical care, vol. 33, pp. 542-551, 2020.

[54] K. H. Goh et al., "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," Nature communications, vol. 12, no. 1, p. 711, 2021.

[55] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, "Combining structured and unstructured data for predictive models: a deep learning approach," BMC medical informatics and decision making, vol. 20, pp. 1-11, 2020.

[56] A. Elmoheen et al., "External validation and recalibration of the CURB-65 and PSI for predicting 30-day mortality and critical care intervention in multiethnic patients with COVID-19," International Journal of Infectious Diseases, vol. 111, pp. 108-116, 2021.

[57] A. S. Darwich et al., "Model-informed precision dosing: background, requirements, validation, implementation, and forward trajectory of individualizing drug therapy," Annual review of pharmacology and toxicology, vol. 61, no. 1, pp. 225-245, 2021.

[58] P. Hanlon et al., "COVID-19–exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study," Wellcome Open Research, vol. 5, 2020.

[59] L. L. Guo et al., "Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine," Scientific reports, vol. 12, no. 1, p. 2726, 2022.

[60] B. Eini-Porat, O. Amir, D. Eytan, and U. Shalit, "Tell me something interesting: Clinical utility of machine learning prediction models in the ICU," Journal of Biomedical Informatics, vol. 132, p. 104107, 2022.

[61] D. Horner, C. Ambrose, C. Taylor, and L. Erinle, "Intensive Care Society State of the Art (SOA) 2023 Congress Abstracts," Journal of the Intensive Care Society, vol. 24, no. 2 Supplement 1, p. 196, 2023.

[62] E. D'Hondt, T. J. Ashby, I. Chakroun, T. Koninckx, and R. Wuyts, "Identifying and evaluating barriers for the implementation of machine learning in the intensive care unit," Communications Medicine, vol. 2, no. 1, p. 162, 2022.

APPENDIX A

TABLE VIII. SELECTED QUANTITATIVE PREDICTORS WITH CORRESPONDING INFORMATION GAIN

Variable	Feature	Item Name	Information Gain	Item ID	Table
X1	Age	Age	0.1555		Admissions
X2	Heart Rate	Heart Rate	0.3602	211 220045	Chartevents
X3	Systolic Blood Pressure	Noninvasive Systolic Blood Pressure	0.2702	455 220179	Chartevents
X4	Temperature	Temperature Fahrenheit Temperature Celsius	0.4681	678 223761 676 223762	Chartevents
X5	Blood Urea Nitrogen	Blood Urea Nitrogen	0.2172	51006	Labevents
X6	White Blood Cells Court	White Blood Cells	0.4725	51301 51300	Labevents
X7	Potassium Level	Potassium	0.2287	50971 50822	Labevents
X8	Sodium Level	Sodium	0.2486	50983	Labevents
X9	Serum Bicarbonate Level	Bicarbonate	0.0954	50882	Labevents
X10	Glasgow Coma Scale	GCS Verbal Verbal Response GCS Motor Motor Response GCS Eyes Eye Opening	0.4340	223900 723 223901 454 220739 184	Chartevents
X11	Gender	Gender	0.0104		Admissions
X12	Admission Type	Admission Type	0.0184		Admissions