International Journal of
Environmental Research and Public Health

MDPI

Article

Integrating Structured and Unstructured EHR Data for Predicting Mortality by Machine Learning and Latent Dirichlet Allocation Method

Chih-Chou Chiu [1], Chung-Min Wu [1], Te-Nien Chien [2,*], Ling-Jing Kao [1], Chengcheng Li [2] and Chuan-Mei Chu [2]

[1] Department of Business Management, National Taipei University of Technology, Taipei 106, Taiwan
[2] College of Management, National Taipei University of Technology, Taipei 106, Taiwan
* Correspondence: tenienchien@gmail.com; Tel.: +886-2-2771-2171 (ext. 3403)

Abstract: An ICU is a critical care unit that provides advanced medical support and continuous monitoring for patients with severe illnesses or injuries. Predicting the mortality rate of ICU patients can not only improve patient outcomes, but also optimize resource allocation. Many studies have attempted to create scoring systems and models that predict the mortality of ICU patients using large amounts of structured clinical data. However, unstructured clinical data recorded during patient admission, such as notes made by physicians, is often overlooked. This study used the MIMIC-III database to predict mortality in ICU patients. In the first part of the study, only eight structured variables were used, including the six basic vital signs, the GCS, and the patient's age at admission. In the second part, unstructured predictor variables were extracted from the initial diagnosis made by physicians when the patients were admitted to the hospital and analyzed using Latent Dirichlet Allocation techniques. The structured and unstructured data were combined using machine learning methods to create a mortality risk prediction model for ICU patients. The results showed that combining structured and unstructured data improved the accuracy of the prediction of clinical outcomes in ICU patients over time. The model achieved an AUROC of 0.88, indicating accurate prediction of patient vital status. Additionally, the model was able to predict patient clinical outcomes over time, successfully identifying important variables. This study demonstrated that a small number of easily collectible structured variables, combined with unstructured data and analyzed using LDA topic modeling, can significantly improve the predictive performance of a mortality risk prediction model for ICU patients. These results suggest that initial clinical observations and diagnoses of ICU patients contain valuable information that can aid ICU medical and nursing staff in making important clinical decisions.

Keywords: structured vs. unstructured data; machine learning; intensive care units; electronic health records; predictive modeling

Citation: Chiu, C.-C.; Wu, C.-M.; Chien, T.-N.; Kao, L.-J.; Li, C.; Chu, C.-M. Integrating Structured and Unstructured EHR Data for Predicting Mortality by Machine Learning and Latent Dirichlet Allocation Method. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4340. https://doi.org/10.3390/ijerph20054340

Academic Editors: Fabio Mendonca, Morgado Dias and Sheikh Shanawaz Mostafa

Received: 16 January 2023
Revised: 22 February 2023
Accepted: 24 February 2023
Published: 28 February 2023

1. Introduction

The World Federation of Societies of Intensive and Critical Care Medicine defines an intensive care unit (ICU) as an organized system of care for critically ill patients that provides intensive and specialized medical and nursing care, enhanced monitoring capabilities, and multiple physiological organ support modality to sustain life during a period of multiple organ dysfunction syndrome (MODS) [1]. The hospital has established an intensive care unit (ICU) for patients with severe or life-threatening conditions. ICU mortality and costs are the highest of all hospital units [2]. It is difficult for medical and nursing staff to deal with rapidly changing patient conditions if there is not enough real-time information for clinicians to make accurate and timely decisions [3]. Different types of judgment errors can have many negative consequences, and incorrect decisions or delayed diagnosis can have a significant impact on patient prognosis, medical resource availability, and healthcare costs [4]. Recently, when the COVID-19 pandemic flooded intensive care units around

the world, their significance was highlighted. In times such as these, more active research on how to manage scarce critical care resources is required to provide additional tools to support medical decision-making and effective clinical practice benchmarks [5]. In the United States, more than 5 million patients are admitted to the ICU annually, and 40% of these patients die during their hospital stay, with 22% spending their entire hospital stay in the ICU [6]. Predicting mortality in ICU patients is one of the most important tasks in critical care research, not only to aid health professionals in clinical decision-making, but also as a basis for managing hospital resource utilization. Patients admitted to the ICU require close and constant monitoring to prevent rapid deterioration of their health. Intensive monitoring through ICU equipment generates a large number of medical records, requiring an efficient and accurate data analysis system [7].

The electronic health record (EHR) is a digital version of the paper chart. Numerous researchers have utilized EHR database data in the past to predict patient mortality, admission time, disease diagnosis, disease onset, etc., to prevent and intervene in early disease in patients which is crucial to critical care. As an essential risk assessment tool, the predictive model has been developed and utilized in numerous healthcare fields. The Sequential Organ Failure Assessment (SOFA), a new Simplified Acute Physiology Score (SAPSII), and the Multiple Organ Dysfunction Score (MODS) have also been used widely in clinical practice to predict mortality [8–10]. Predictive models facilitate the early identification of patients at risk for a disease or event and provide effective intervention measures for those who are most likely to benefit from the identification of specific risk factors. Much research has been conducted to determine how data analysis and prediction can assist medical and nursing staff in the process of diagnosis and treatment to heighten alertness to the progression of patient condition [11–15]. The results of the statistical data's predictive power derived from the basic vital signs and simple demographic data such as age save the most resources and are the most useful. Vital signs were chosen as features mainly because most vital signs can be easily measured using non-invasive equipment, and vital signs are the most basic health indicators that are easily understood by all healthcare professionals [16–19].

Unstructured data comprise 80% of EHR data [20]. It is undeniable that overlooking the deficiencies of qualitative data in the EHR may not only result in the omission of key factors caused by the absence of handwritten diagnostic data, but may also result in the omission of clues in the initial judgement being overlooked or diminished. Although these variables can be used to partially predict the mortality of ICU patients, quantitative variables are utilized in the majority of these studies. After all, the existing statistical predictive modeling is relatively mature with respect to the processing of quantitative data, whereas distinctive challenges exist in the standardization and utilization of qualitative data [21–23].

Machine learning is a subfield of artificial intelligence concerned with teaching computers to learn from data and improve with experience. It focuses on the issue of how to design computer programs that can automatically improve output accuracy based on experience [24]. There has recently been an increase in the use of machine learning applications in clinical medicine. These include preclinical data processing, bedside diagnostic assistance, patient stratification, treatment decision-making, and early warning for primary and secondary prevention [25]. Machine learning can improve clinical decision-making in many ways, by providing early warning, facilitating diagnosis, conducting widespread screening, personalizing treatment, and assessing patient response to treatment. Many different fields and clinical applications are gradually adopting machine learning from mature preclinical scenarios [26,27].

The development of machine learning includes text mining, natural language processing and Latent Dirichlet Allocation (LDA) which are used to identify and extract information or relationships from unstructured data and have become popular techniques for literary analysis [28,29]. LDA is a Bayesian probability generation model in the field of natural language processing proposed by Blei et al. [30] that has several advantages for literature analysis. LDA is a powerful tool for processing massive amounts of data

that can capture text-specific dimensions without relying on assumptions. Furthermore, it incorporates multiple steps of text analysis, such as data sampling with minimal human intervention to yield more realistic and objective topic modeling outcomes [31,32].

However, it takes a lot of time and money to process the unstructured data that make up medical big data. This is particularly so for the digital part, typically a vital component, presented in the large number of clinical notes made during treatment and hospital stay. In accordance with the rules of unstructured data processing, numbers are frequently removed to reduce their utility. By incorporating unstructured data as input, we are not using raw physiological data, but rather the perception and judgment of medical and nursing professionals in the form of free text annotations. These allow us to access higher-level concepts that are not present in the physiological data. The text data format is relatively consistent, and this allows circumvention of the LDA digital deletion limitation. This is the most noticeable feature of free text records, which contain information about patients' admission to and diagnosis in the ICU. Data about observations and first signs of condition and diagnoses are added as soon as possible after admission of the patient to the ICU, with minimal interference from the earlier patient data. Clinicians can also use the topic obtained as a follow-up reference. Our recent study combined 16 structured variables and 10 topic modeling semi-structured variables from the Medical Information Mart for Intensive Care (MIMIC-III) dataset to predict mortality in ICU patients. The results show that semi-structured data contain useful information that can help clinicians make critical clinical decisions [33].

In this study, we utilized the MIMIC-III database to develop a model for predicting mortality in ICU patients. Our approach involved integrating structured data, which are basic and easily collected from ICU patients, with unstructured data derived from the initial clinical diagnosis of the patient's physician at the time of admission. We used the LDA approach to topic modeling of diagnostic records and applied machine learning techniques to combine both structured and unstructured data to build a robust mortality risk prediction model. This model can provide patients, their families, and healthcare professionals with valuable additional information for making informed medical decisions. Our findings could have significant implications for improving patient outcomes and advancing critical care medicine.

2. Materials and Methods

2.1. Proposed Framework

Figure 1 depicts the framework of this study. The structured data collected after patients were admitted to the ICU was integrated (six vital sign measurements in the first 24 h, the Glasgow Coma Scale (GCS), and patient age) with unstructured data (initial clinical diagnosis records at ICU admission). The machine learning model was used to predict mortality of the ICU patient. Finally, five different metrics were used to assess predictive performance. The period of mortality in ICU patients is defined as follows:

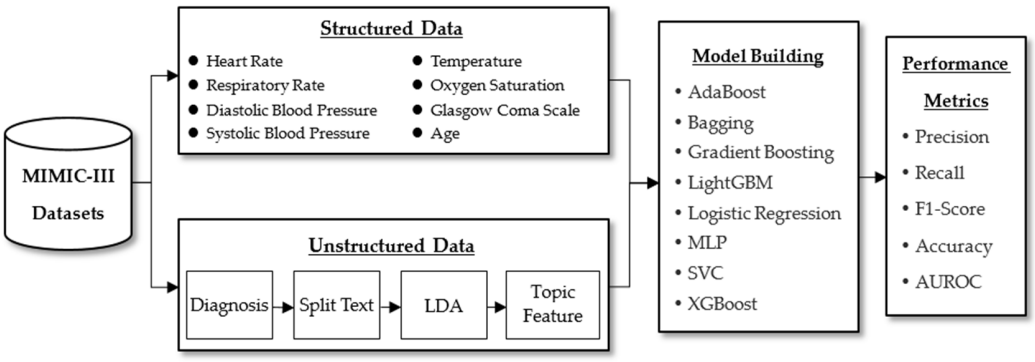

Figure 1. Research scheme.

2.2. Data Collection and Preprocessing

The rapid development of digital health systems has occurred in recent years. However, concerns surrounding personal privacy and security have made it difficult to integrate and apply this information to scientific research. To ensure the convenience and completeness of data collection, this study has focused on obtaining complete patient dynamic information from databases that are easier to obtain than ICU data. The data were obtained from MIMIC-III clinical database in our research. MIMIC-III uses integrated comprehensive clinical data from patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts [34]. The MIMIC-III database used contained information on 46,520 patients and 58,976 admission-related data items, including patient vital signs, drugs, laboratory measurement values, and observation records. There were 38,597 adult patients, 56% of whom were male, and the median age was 65.8. The median of the length of admission was 6.9 days, the mortality rate during admission was 11.5%, and the median of the length of stay in the ICU was 2.1 days. Furthermore, the following data were generated per patient per stay in the intensive care unit: 6643 patient observation records, 83 patient medical document records, and 559 laboratory test result records. Table 1 shows the database compilation. The National Institutes of Health (NIH) online course was completed, as well as an exam for protecting human research participants and the submission of an access application (Certification Number: 35628530).

Table 1. The summary of MIMIC-III dataset.

Distinct Patients	**46,520**
Age, years, median [Q1–Q3]	65.8 [52.8–77.8]
Gender, male, percent of unit stays	26,121 (56.15%)
Distinct hospital admissions	58,976
Elective	7706 (13.07%)
Emergency	42,071 (71.34%)
Newborn	7863 (13.33%)
Urgent	1336 (2.27%)
Hospital mortality, percent of unit stays	5854 (9.93%)
Hospital length of stay, median days [Q1–Q3]	10.13 [3.74–11.80]
Distinct ICU stays	61,532
Coronary Care Unit	7726 (12.56%)
Cardiac Surgery Recovery Unit	9312 (15.13%)
Medical Intensive Care Unit	21,088 (34.27%)
Neonatal Intensive Care Unit	8100 (13.16%)
Surgical Intensive Care Unit	8891 (14.45%)
Trauma Surgical Intensive Care Unit	6415 (10.43%)
ICU length of stay, median days [Q1–Q3]	4.92 [1.11–4.48]

To reflect the universality of the analytical results and to ensure that they were comparable with the conclusions of the related literature, this study followed the patient selection principles of previous related studies and specific diseases in patients were not analyzed; instead, the data from all patients were used [18,32,35]. First, the inclusion of only the initial ICU admission and exclusion of all subsequent ICU readmissions ensured that the outcome was measured the same way for all patients. This highlighted the early predictive ability of the model and prevented possible information omissions when the dataset was separated for training and testing (12,456 admissions were deleted). Second, the subjects used in this study were all adults older than 16 years (7878 admissions were deleted). Lastly, only data from patients who stayed in the ICU for longer than 24 h were utilized (2138 admissions were deleted). In the case of patients who stayed in the ICU for at least one day, only data from their first day were considered. If multiple measurements had been taken on the same day, an average of the values was taken. In addition, the data preprocessing method of Guo et al. [36] was used for the processing of missing values in this study. Three-stage

missing value processing was carried out and patients with more than 30% missing variable values were excluded (8954 admissions were deleted) and a total of 27,550 participants were included in this study. Figure 2 depicts the extraction of data in their entirety.

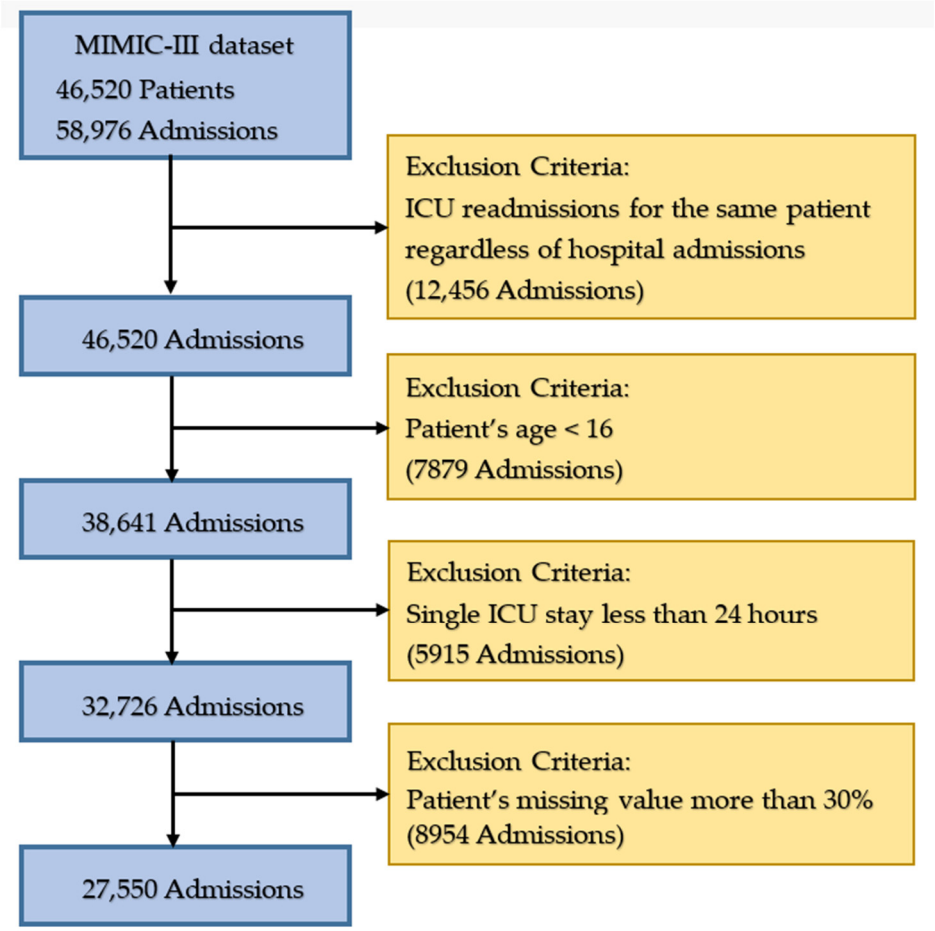

Figure 2. The process of data extraction.

In this study, information from the MIMIC-III database admission and chart events tables were used for the variable selection part. In reference to previous related studies [16–18,37], only six basic vital signs from the patient files were used: Heart Rate, Respiratory Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Temperature, and Oxygen saturation, along with Glasgow Coma Scale and the patient's age at admission as a predictor of variables in the first part. Topic model variables extracted from unstructured data of the initial diagnosis made by physicians when the patients were admitted to the hospital, were among the predictive variables in the second part.

2.3. Baseline Characteristics

Ultimately, the ICU records of 27,550 patients were utilized after the data in the MIMIC-III database had been preprocessed. Table 2 shows patient demographic information. The average age of the patients in this study was 64, of which 56% were male, their average hospital stay was 10.39 days, and their average intensive care unit stay was 4.48 days. In addition, over 84% of patients were admitted to the hospital for emergency care. Medicare insurance covered more than 50% of patients. Table 2 also displays the statistical values of the eight structural variables utilized in the study.

Table 2. Features involved in the model.

	Total	Survivors	Non-Survivors
General			
Number	27,550 (100%)	24,364 (88.44%)	3186 (11.56%)
Gender (male)	15,441 (56.05%)	13,764 (89.14%)	1677 (10.86%)
Length of stay			
Hospital (days) [Q1–Q3]	10.39 [4.36–12.42]	10.29 [4.44–12.35]	11.20 [3.88–12.54]
ICU (days) [Q1–Q3]	4.48 [1.55–4.61]	4.18 [1.53–4.33]	6.77 [1.70–6.13]
Admission Type			
Elective	3537 (12.84%)	3434 (14.09%)	103 (3.23%)
Emergency	23,283 (84.51%)	20,293 (83.29%)	2990 (93.85%)
Urgent	730 (2.65%)	637 (2.61%)	93 (2.92%)
Insurance			
Government	844 (3.06%)	799 (3.28%)	45 (1.41%)
Medicaid	2301 (8.35%)	2078 (8.52%)	223 (7.00%)
Medicare	14,750 (53.54%)	12,618 (51.79%)	2132 (66.92%)
Private	9303 (33.77%)	8570 (35.17%)	733 (23.01%)
Self-Pay	352 (1.28%)	300 (1.23%)	52 (1.63%)
Variable value (First 24 h)			
Heart Rate	85.72 ± 15.91	85.11 ± 15.52	90.38 ± 17.93
Respiratory Rate	18.94 ± 4.01	18.70 ± 3.84	20.75 ± 4.76
Diastolic Blood Pressure	60.33 ± 12.21	60.66 ± 12.09	57.79 ± 12.79
Systolic Blood Pressure	118.01 ± 18.60	118.45 ± 18.30	114.63 ± 20.45
Temperature	98.23 ± 2.03	98.28 ± 1.73	97.90 ± 3.54
Oxygen Saturation	97.21 ± 2.11	97.28 ± 1.90	96.68 ± 3.28
Glasgow Coma Scale	12.31 ± 3.21	12.66 ± 2.91	9.64 ± 4.00
Age	64.00 ± 17.71	63.13 ± 17.75	70.66 ± 15.81

Among these were the diagnosis in the initial clinical notes about the patient made by the physician. As shown in Table 3, the diagnosis field provides the clinician with a written record of the initial diagnosis on admission. The admitting clinician usually specifies a diagnosis and does not use system ontology. Diagnoses may be very useful (e.g., congestive heart failure\biventricular implantable cardioverter defibrillator placement) or extremely vague (e.g., fever). This text section can provide useful information about the condition of the patient on admission. The information in the diagnosis field from the Admission Table was used in this study and a machine learning model was used to investigate the impact of structured EHR data and unstructured data on ICU patient mortality. Structured EHR data included variables such as vital signs and lab tests, and clinical note content includes topic features extracted from clinical notes using the LDA method.

Table 3. Patients' diagnosis records.

SUBJECT_ID	HADM_ID	Diagnosis
00412	109897	AORTIC STENOSIS; MITRAL REGURGITATION; CAD\AORTIC VALVE REPLACEMENT; MITRAL VALVE REPLACEMENT; CORONARY ARTERY BYPASS GRAFT; TRICUSPID VALVE REPLACEMENT/SDA
00969	137250	BATTERY DEPLETION; HEART FAILURE\IMPLANTABLE CARDIOVERTER DEFIBRILLATOR EXPLANT; PACEMAKER IMPLANT; DIURISIS POST PROCEDURE/SDA
14229	145873	MARFAN'S SYNDROME\BENTALL PROCEDURE; TOTAL VALVE SPARING ROOT REPLACEMENT VS; HOMOGRAFT ROOT REPLACEMENT; REPLACEMENT OF ARCH, PROXIMAL ROOT/SDA
22416	130625	DESCENDING AORTIC ANEURYSM; COARCTATION OF DESCENDING AORTA\ DISTAL ARCH REPLACEMENT; DESCENDING THORACIC AORTIC REPLACEMENT; AORTA TO SUBCLAVIAN BYPASS/SDA
23360	104836	POLYCHONDRITIS WITH AIRWAY MANISFESTATION\ STERNATOMY CARDIOPULMONARY; BYPASS; ANTERIOR TRACHEAL SPLITTING; TY STENT PLACEMENT; LAPAROTOMY/SDA

Table 3. *Cont.*

SUBJECT_ID	HADM_ID	Diagnosis
28352	154475	PULMONARY VEIN INJURY\THORACOSCOPIC MAZE PROCEDURE LEFT; MINI MAZE; BILATERAL MINI THORACOTOMY; PULMONARY VEIN ISOLATION; RESECTION OF LEFT ATRIAL APPENDAGE/SDA
45688	144761	RIGHT VENTRICULAR LEAD MALFUNCTION; INAPPROPRIATE IMPLANTABLE CARDIOVERTER-DEFIBRILLATOR FRING\RIGHT VENTRICULAR IMPLANTABLE CARDIOVERTER-DEFIBRILLATOR LEAD EXTRACTION/SDA
51821	182983	MEDIASTINAL ADENOPATHY\FLEXIBLE BRONCHOSCOPY; LINEAR ENDOBRONCHIAL ULTRASOUND (EBUS); FLUOROSCOPY; TRANSBRONCHIAL BIOPSY; TRANSBRONCHIAL NEEDLE ASPIRATION; BRONCHIAL ALVEOLAR LAVARGE
92284	193856	AIRWAY OBSTRUCTION\FLEXIBLE BRONCHOSCOPY; RADIAL ENDOBRONCHIAL ULTRASOUND (EBUS); BRONCHIAL AVEOLAR LAVAGE/ BRUSH; POSSIBLE TRANSBRONCHIAL BIOPSY (LEFT UPPER LOBE); FLUOROSCOPY

2.4. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised topic modeling algorithm that derives topics in a corpus. The model is a standard "bag of words" model, wherein each text item is viewed as a word frequency vector and the text is viewed as a set made up of various word groups [30]. Typically, an LDA topic generation model is built in three steps: First, a topic is extracted from the topic distribution for each text item. Second, a vocabulary corresponding to the extracted topics is taken from the vocabulary distribution. The steps are then repeated until every word in the text has been extracted. Because each text item contains multiple topics, several corresponding key words can be chosen for each topic. In other words, the same vocabulary can appear across multiple topics. Topic modeling methods mine significant topics from collected documents using probabilistic procedures and applications. As a result, by effectively processing a large amount of unstructured data in the text, the topic modeling method can help identify the latent semantics of complex articles [38,39]. LDA assumes that each document in the collection is created in two steps, the first by selecting a distribution of topics for that document, and the second by assigning a random topic and its corresponding distribution of words to each position in the document that may contain a word. This is repeated for the entire corpus. As a result, the main feature of LDA is that all documents share the same topic to varying degrees. Based on this theory, an LDA model can be applied to a set of documents using the Gibbs sampling algorithm to infer their underlying topics. The algorithm iterates over all the words in the document and calculates the most representative words for each topic. Each word can appear multiple times in the same document and can be repeated in different documents at the same time. At each iteration, the algorithm can modify the topic that best represents it, and after using Gibbs sampling with the training set, a model is built that produces a topic distribution for each document [40].

In a similar context-based textual analysis, probabilistic topic modeling conceptualizes a document as a collection of words derived from underlying thematic topics that define a probability distribution of words related to a topic, where the relative importance of each word in respective topics is defined by the conditional probability $P(Word_i|Topic_j)$ in category probability distribution. Because an article is a weighted mixture of multiple topics, its conditional probability can be determined, and file content is generated based on the proportion of words related to each topic. This matrix is decomposed by topic modeling approaches based on latent topic structures that link latent words to related documents. A precise solution to this inverted inference is not generally tractable and requires an iterative optimization solution such as that given by Gibbs sampling. The probabilistic LDA framework will interpret correlation structures as conditional probabilities $P(Word_i|Topic_j)$ and $P\left(Topic_j|Document_k\right)$, which are closely related to other dimensionality reduction techniques for providing low-rank data approximations. An insight into the underlying

topic structure allows for a more convenient, efficient, and interpretable approach to information retrieval, classification, and document data exploration [41].

$$P(\text{Word}_i \,|\text{Document}_k) = \sum_{j=1}^{J} P\left(\text{Word}_i \,|\text{Topic}_j\right) \times P\left(\text{Topic}_j \,|\text{Document}_k\right) \quad (1)$$

In the medical field, topic modeling research primarily focuses on the organization of clinical text, such as in newspapers and scientific literature, as well as clinical discharge records. However, recent studies have modeled laboratory results, claims data, and clinical concepts [42,43]. In this study, the aim is to learn the topic structure of clinical data through algorithms and apply it to clinical decision-making prediction. Unlike a top-down rule-based approach that isolates preconceived clinical concepts from electronic medical records, this bottom-up approach recognizes patterns in the data with more consistency. Additionally, in this paper, reference is made to an algorithm used in a previous study for the handling of non-quantified data [44,45], where Grid Search is used to confirm the best LDA model and tests multiple sets of topics. The LDA model was then applied to the ten topics derived from the results to categorize these key words into different topics. Any word that appears in a keyword set is related to the topic. Furthermore, certain words are more likely to appear under each topic, and there is a probability that each word will appear under respective topics. Figure 3 and Table 4 show the ten topics and keywords chosen for topic modeling in this study.

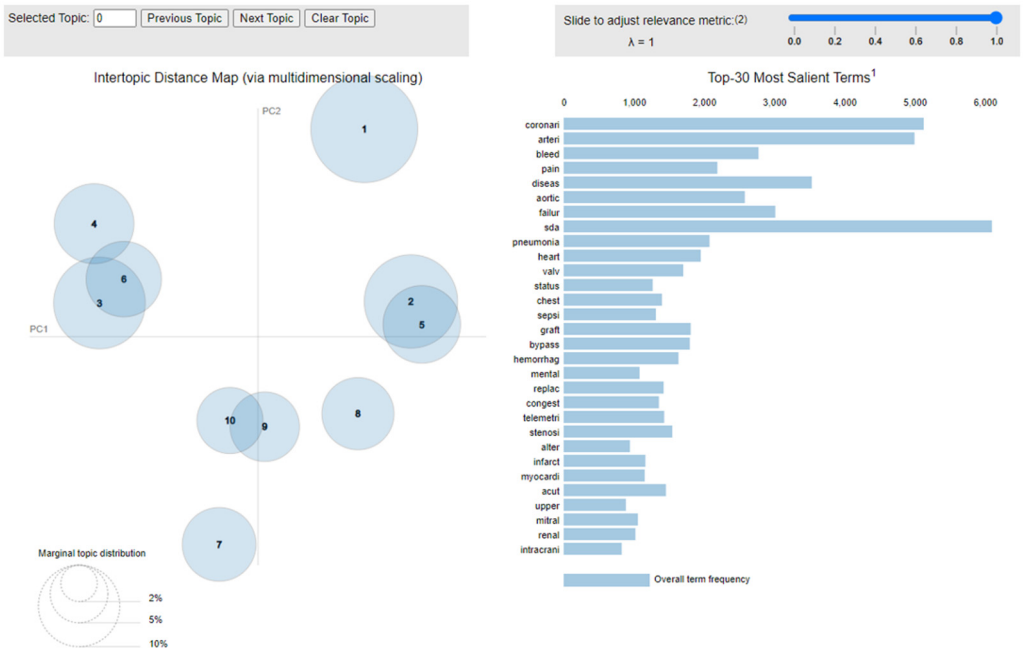

Figure 3. Topic modelling in Python.

Table 4. Topics and keywords for dataset.

Variable	Topic	Keywords
$TOPIC_1$	Coronary Artery Disease	coronari, arteri, diseas, graft, bypass, sda, syndrom, effus, cath, avr, acut, etoh, pericardi, cerebr, pleural, cholang, mvr, leav, vascular, angioplasty.
$TOPIC_2$	Aortic Valve Replacement	aortic, sda, valv, replac, stenosi, mitral, cancer, subarachnoid, hemorrhag, esophag, procedur, ascend, regurgit, aorta, maze, airway, redo, bental, repair, invas, obstruct
$TOPIC_3$	Heart Failure	failur, heart, congest, acut, infarct, myocardi, renal, cath, liver, pancreat, elev, cardiac, dehydr, rule, cholecyst, hyperkalemia, leukemia, lacer, hyperglycemia, block, chronic, implant
$TOPIC_4$	Pneumonia	pneumonia, telemetri, fractur, stroke, ischem, atrial, attack, transient, angina, dyspnea, fibril, hip, unstabl, cath, chronic, segment, diseas, pulm, obst, ablat, cardiomyopathi, pelvic, septal

Table 4. *Cont.*

Variable	Topic	Keywords
$TOPIC_5$	SDA	sda, right, aneurysm, leav, accid, motor, vehicl, short, breath, tachycardia, cellul, ventricular, abdomin, lung, injuri, hepat, bilater, metastat, perfor, colon, spinal
$TOPIC_6$	Chest Pain	pain, chest, hemorrhag, intracrani, fever, abdomin, hypotens, fall, cath, telemetri, dissect, insuffici, stroke, cardiac, femur, strike, epidur, neck, pedestrian, skull, cervic
$TOPIC_7$	Bleed Mass	bleed, upper, lower, mass, pulmonari, obstruct, bowel, head, brain, weak, emboli, bradycardia, hypertens, stemi, small, edema, hemoptysi, cirrhosi, vomit
$TOPIC_8$	Sepsis	sepsi, infect, asthma, exacerb, copd, sda, urinari, tract, tumor, brain, catheter, overdos, leav, anemia, pyelonephr, syncop, bscess, foot, ulcer, disord
$TOPIC_9$	Subdural Hematoma	hematoma, subdur, respiratori, diabet, seizur, ketoacidosi, trauma, failur, blunt, sda, withdraw, distress, hyponatremia, wind, hernia, remot
$TOPIC_{10}$	Altered Mental Status	status, mental, alter, arrest, cardiac, carotid, hypoxia, chang, leg, angiogram, stenosi, transplant, kidney, chf, accid, extrem, cerebrovascular, fib, stent, thrombosi, ischemia

2.5. Machine Learning

In this study, the organized dataset was divided into two parts with 80% of the data being used for training the model and the remaining 20% for testing. Eight commonly used machine learning algorithms were used to establish the ICU mortality prediction model: Adaptive Boosting (AdaBoost), Bagging, Gradient Boosting, Light Gradient Boosting Machine (LightGBM), Logistic Regression, Multilayer Perceptron (MLP), Support Vector Classification (SVC), eXtreme Gradient Boosting (XGBoost). All data mining tasks of this research were conducted using the Python programming language. Table 5 shows the 8 machine learning models with their specific parameters' settings. The following sections provide detailed descriptions of the various machine learning classification algorithms.

Table 5. Machine learning models with their specific parameters' settings.

Model	Parameters
AdaBoost	*base_estimator = DecistionTreeClassifer, random_state = 1, n_estimators = 50, learning_rate = 1.0, algorithm = 'SAMME.R'*
Bagging	*base_estimator = None, n_estimators = 500, max_samples = 100, bootstrap = True, bootstrap_features = False, oob_score = False, warm_start = False, n_jobs = 1, random_state = None, verbose = 0*
Gradient Boosting	*n_estimators = 100, learning_rate = 1.0, max_depth = 1, random_state=0*
LightGBM	*boosting_type = 'gbdt', num_leaves = 31, max_depth = −1, learning_rate = 0.1, n_estimators = 100, subsample_for_bin = 200,000, min_child_samples = 20, subsample = 1.0, subsample_freq = 0, colsample_bytree = 1.0, reg_alpha = 0.0, reg_lambda = 0.0 n_jobs = −1, importance_type = 'split',*
Logistic Regression	*solver = 'sag', penalty = 'l2', max_iter = 'max_iter'*
MLP	*solver = 'adam', alpha = 1e-5, hidden_layer_sizes = (13,13,13), max_iter = 1000*
SVC	*C = 1.0, kernel = 'rbf', degree = 3, gamma = 'auto', coef0 = 0.0, shrinking = True, probability = False, tol = 0.001, cache_size = 200, class_weight = None, verbose = False, max_iter = −1*
XGBoost	*n_estimators = 100, booster = 'gbtree', eta = 0.3, min_child_weight = 1, max_depth = 3, gamma = 0, max_delta_step = 0, su-bsample = 1, colsample_bytree = 1, colsample_byleve = 1, lambda = 1, learning_rate = 0.1, n_jobs = 1, base_score = 0.5, max_delta_step = 0, min_child_weight = 1*

- AdaBoost is an adaptive method in the sense that incorrect samples from the previous classifier are used to train the next classifier. The AdaBoost method is sensitive to noise and abnormal data. It trains a basic classifier and gives misclassified samples more weight. It is then applied to the next process. This iterative process is repeated until a stopping condition is reached or the error rate is low enough [46,47]. The Python sklearn library was used to implement AdaBoost. Our hyperparameters specified a

maximum number of iterations of 50, while others trained the model using the sklearn preset values.

- The Bootstrap Aggregating algorithm, also known as the Bagging algorithm, is an ensemble learning algorithm in the field of machine learning, which was first proposed by Leo Breiman in 1994. The Bagging algorithm can be combined with other classification and regression algorithms to improve accuracy and stability while reducing result variance to avoid overfitting. Bagging is an ensemble method that combines multiple predictors. It helps to prevent model overfitting to data and reduces variance. It has been used in many microarray studies [48,49]. The Python sklearn library was also used to implement bagging. A combined classifier made up of 500 DecisionTreeClassifiers was used. Each classifier has a maximum sampling subset of 100; the self-service sampling method was used for each sampling. Other hyperparameters used sklearn preset values to carry out training.
- Gradient Boosting is an ensemble learning algorithm that can be used to improve the accuracy of various classification prediction models. It trains a model with poor prediction accuracy using the negative gradient information of the model loss function and then combines the trained results with the existing model in a cumulative form [50]. The scikit-learn library was also used in this study to achieve gradient boosting; the maximum number of iterations was set to 100; and other hyperparameters were trained using preset scikit-learn library values.
- The Light Gradient Boosting Machine (LightGBM) is an ensemble method that combines the predictions of multiple decision trees to produce a well-generalized final prediction. LightGBM divides continuous eigenvalues into K intervals and chooses dividing points from those intervals. This method significantly accelerates prediction and reduces memory occupancy without sacrificing prediction accuracy [51]. LightGBM is a decision tree learning algorithm with gradient boosting that has been widely used for feature selection, classification, and regression [52].
- Logistic Regression is a logit model capable of testing statistical interactions and controlling multivariate confidence. It is most commonly used to investigate the risk relationship between disease and exposure [53,54]. In this study, the Python scikit-learn library was used to implement logistic regression, and the hyperparameter optimization method was the SAG linear convergence algorithm. It is a gradient descent method used specifically for large sample data.
- Multilayer Perceptron (MLP) is a feed-forward artificial neural network with a fixed number of computational units or neurons that are fully connected to the next layer [55]. A multilayer perceptron learns and predicts data using the principles of the human nervous system. MLPs are suitable for classifying and predicting tasks with different feature set implementations [56]. The neural network used in this study had five layers: an input layer, three hidden layers, and an output layer. Each hidden layer has 13 neuron nodes, the normalization parameter was set to 1e-5, relu was used as the activation function, and adam was used for training and weight optimization.
- The Support Vector Classifier (SVC) analyzes linear and nonlinear data for classification and regression. SVC aims to recognize categories by the creation of non-linear decision hyperplanes in a higher feature space [57]. SVC is resistant to data bias and variance and produces accurate predictions for binary or multiclass classifications. Additionally, SVC is robust, resists overfitting, and has exceptional generalization capabilities [58].
- eXtreme Gradient Boosting (XGBoost) is a scalable end-to-end tree boosting system that is an optimized implementation of the gradient boosting framework. It is remarkable in that it can handle missing data efficiently, is very flexible, and can build an assembly of weak prediction models into an accurate one [59]. It generates a series of decision trees during training, each building on the previous one to reduce the loss function gradient. Furthermore, a predictive model made up of multiple decision trees can be obtained. The XGBoost algorithm can deal with missing values by includ-

ing a default orientation for missing values in each tree node and learning the best orientation from the data [60].

2.6. The Synthetic Minority Oversampling Technique (SMOTE)

When the class distribution is highly skewed, machine learning problems become unbalanced. Unbalanced classification problems are prevalent in a variety of application domains and pose challenges for conventional learning algorithms [61]. In general, an imbalanced dataset can negatively affect the results of a model. In general, gold-standard datasets are unbalanced, which reduces model predictive ability [62]. In the evaluation of model performance, over- and underfitting are the most common issues. When a model has a high accuracy score during training but a low accuracy one during verification, overfitting has occurred. The greatest reduction in model overfitting can be achieved by increasing the size of the training set and decreasing the number of neural network layers. The failure of a model to classify data or make predictions during the training phase signifies underfitting [63]. SMOTE is a potent classification imbalance solution that produces consistent results across domains. The SMOTE algorithm adds synthetic data to the minority class to create a balanced dataset [61]. Class imbalance refers to the disparity between the classes of data used to train a predictive model, a prevalent issue that is not exclusive to medical data. Classification algorithms have a tendency to favor the majority class when it has significantly fewer observations than the class with negative outcome. Predictive performance can be improved by the manipulation of data, algorithms, or both [64]. The methodology involves the under- and oversampling of larger and smaller samples.

Table 6 displays the descriptive statistics for the data used in this study. The data in the table indicate a significant imbalance between the ratio of patient survival and mortality. Because these unbalanced datasets frequently produce inaccurate model prediction [65], the addition of minority class samples, or the deletion of majority class samples, is frequently performed to correct this [15]. The Synthetic Minority Oversampling Technique (SMOTE) randomly generates new minority class samples from the nearest neighbor line connecting the minority class samples and the technique is extensively used to process skewed data [63,66]. In this study, SMOTE technology was used to increase the sample size for the side with fewer samples to balance the data [15]. This was necessary because the number of samples of patients dying in the ICU was much smaller than the number of samples of patients surviving. In other words, a synthetic minority sampling technique was used to preprocess extremely unbalanced datasets.

Table 6. Demographic information and SMOTE technique of the selected patient cohort.

	3 Days	30 Days	365 Days
Number of patients	27,550	27,550	27,550
Number of survivors	26,640	24,522	24,364
Number of non-survivors	910	3028	3186
Mortality ratio	3.30%	10.99%	11.56%
SMOTE increase	2900%	900%	900%
Number of survivors	26,640	24,522	24,364
Number of non-survivors	26,390	24,224	22,302
Mortality ratio	49.76%	49.69%	47.79%

In this study, a range of SMOTE methods of varying percentages was used to examine a selection of cases. A fresh training dataset was produced based on the information in Table 6. Non-survivors' samples were increased by a factor of eight or nine using the SMOTE technology on a dataset of patients who died within 30 days of admission, from 3028 patients to 24,224 patients. This increased the proportion of the minority group in the baseline dataset from 10.99% to 49.69%.

2.7. Performance Evaluation

To make a thorough comparison of the impact of the integration of structured and unstructured data on the prediction of mortality in ICU patients, in this study, five different metrics were chosen as evaluation tools for modeling. These included AUROC, Precision, Recall, F1-Score, and Accuracy. Appendix A shows the confusion matrix.

3. Results

3.1. Prediction of Mortality in ICU

The k-fold cross-validation method was used to assess the performance of the model after training. The dataset was initially divided into k sections, with each section containing instances of equal size. The final measure of performance was the average of all test results across all components. This method has the benefit of training and validating all instances of the entire dataset, resulting in more accurate predictions with less bias. However, it is computationally costly, and validation is time-consuming. The model was constructed using 10-fold cross-validation, which has been utilized in a number of healthcare and medical studies [67,68]. In this study, the patient's mortality was predicted at 3 days, 30 days, and 365 days after admission based on data collected within 24 h of admission. AUROC, which compares the true-positive rate to the false-positive rate, is the most prevalent metric used to evaluate the performance of diagnostic tools. Table 7 lists the eight distinct machine learning methods employed in this study, as well as AUROC for the ICU mortality prediction task across three all time periods. Our AUROC findings revealed that the mortality rate in 3 days can exceed 80%, and in 30 days and 365 days can exceed 75%. The results indicated that the best AUROC is 88.20% in our research, and that could accurately predict patient death within 3 days at 24 h after admission. Compared with using structured quantitative data alone, adding unstructured data makes the model increase by 2–5% on average in AUROC, which is a great improvement in the prediction of mortality of patients in intensive care units. Figure 4 shows that ICU data can be used to predict 3-day mortality with better precision. This clearly shows that the model developed in this study can predict the vital status of patients with great precision. Gradient Boosting, as indicated by the data in the chart, is the best model for predicting ICU patient mortality across all time periods.

Table 7. AUROC of different classifiers.

		3 Days	30 Days	365 Days
Structured Data	AdaBoost	0.8530 ± 0.0041	0.7514 ± 0.0095	0.7478 ± 0.0058
	Bagging	0.8568 ± 0.0073	0.7627 ± 0.0054	0.7526 ± 0.0049
	Gradient Boosting	0.8598 ± 0.0082	0.7634 ± 0.0126	0.7588 ± 0.0053
	LightGBM	0.8159 ± 0.0149	0.7594 ± 0.0062	0.7523 ± 0.0035
	Logistic Regression	0.8110 ± 0.0221	0.7396 ± 0.0076	0.7353 ± 0.0063
	MLP	0.8494 ± 0.0163	0.7571 ± 0.0097	0.7493 ± 0.0060
	SVC	0.8097 ± 0.0040	0.7487 ± 0.0103	0.7443 ± 0.0056
	XGBoost	0.7070 ± 0.0115	0.7215 ± 0.0070	0.7201 ± 0.0041
Structured Data + Unstructured Data	AdaBoost	0.8686 ± 0.0076	0.7531 ± 0.0066	0.7629 ± 0.0114
	Bagging	0.8713 ± 0.0059	0.7644 ± 0.0098	0.7725 ± 0.0048
	Gradient Boosting	0.8820 ± 0.0119	0.7815 ± 0.0073	0.7754 ± 0.0091
	LightGBM	0.8361 ± 0.0143	0.7780 ± 0.0118	0.7705 ± 0.0036
	Logistic Regression	0.8298 ± 0.0109	0.7618 ± 0.0102	0.7502 ± 0.0051
	MLP	0.8679 ± 0.0168	0.7693 ± 0.0112	0.7540 ± 0.0042
	SVC	0.8142 ± 0.0082	0.7518 ± 0.0110	0.7512 ± 0.0040
	XGBoost	0.7655 ± 0.0174	0.7379 ± 0.0067	0.7345 ± 0.0044

For a comprehensive understanding of the impact of unstructured data on the prediction of mortality in ICU patients, the prediction results of models using only structured data within 24 h of ICU patient admission were compared with those using both structured

and unstructured data. As illustrated in Figure 4, the pertinent prediction results are sorted. Within 24 h of ICU patient admission, the ROC of model prediction results using both structured and unstructured data is greater than that predicted using only structured data across all time periods. Table 7 also demonstrates that Gradient Boosting has a higher AUROC than other machine learning algorithms, regardless of ICU patient mortality across all time periods. Moreover, the prediction accuracy of the model made using both structured and unstructured data, within 24 h of patient admission to the ICU, is generally higher than that of the model using structured data alone. Indeed, basic observations and judgments of the patient at the time are of reference value and will significantly influence the accuracy of constructed model predictions. Overall, this indicates that a model constructed using both structured and unstructured data from ICU patients after admission can predict early patient death after admission with considerable accuracy. By incorporating unstructured data as input, it is possible to gain access to higher-level concepts not present in physiological data.

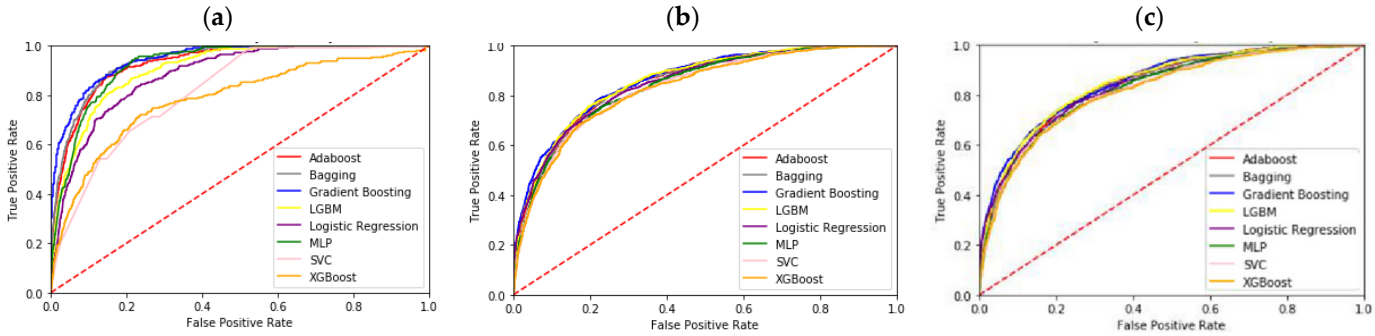

Figure 4. ROC curves for the different classifiers. (**a**) 3-day mortality. (**b**) 30-day mortality. (**c**) 365-day mortality.

We summarize the use of four different metrics (Precision, Recall, F1-Score, and Accuracy) for a more complete picture of the differences in prediction accuracy of models constructed using different machine learning methods in Appendix B, which also show an evaluation of the structured ICU patient basic vital signs within 24 h of admission. These basic observations and judgement depend on whether or not the model was made using unstructured data about patient condition collected at time of admission to the ICU to predict time of patient death. According to the data in the table, the results obtained by using both structured and unstructured data of ICU patients after admission (and the eight different machine learning methods) are slightly better at predicting ICU patient mortality than those using structured data alone under different evaluation metrics. Furthermore, XGBoost has the highest prediction accuracy (97.23%) of the algorithms used, followed by LightGBM (95.61%), and Bagging has the highest prediction recall (95.13%)

3.2. Feature Importance

The most promising features are typically chosen, and the unimportant ones are usually eliminated using feature selection methods. The feature importance score reflects the information gained by each feature during construction of the decision tree [69]. An advantage of using Gradient Boosting is that, once the prediction model has been constructed, the variable importance can be obtained with relative ease by sorting the calculated variable importance scores. The feature importance framework ranks input variables according to their contribution to the predictive model and gives insight into which features are crucial for the task [70]. The more a variable is utilized in the decision tree, the more important it will become. In this study, the importance of each feature is determined by applying a feature importance scoring method to a model trained with gradient boosting. In addition, a percentage rating is provided for how frequently each feature is used to determine the output label. Relevant research notes [71,72] provide additional information on how the

Gradient Boosting method determines the significance of input variables. The variance importance within 24 h of ICU patient admission is outlined in Table 8.

Table 8. The important variables by using Gradient Boosting.

Dataset	Variable Importance	3 Days	30 Days	365 Days
Structured Data	1	X_7	X_7	X_7
	2	X_8	X_8	X_8
	3	X_1	X_1	X_1
	4	X_4	X_4	X_4
	5	X_3	X_5	X_5
Structured Data + Unstructured Data	1	$TOPIC_6$	X_7	X_7
	2	X_7	$TOPIC_1$	$TOPIC_1$
	3	$TOPIC_1$	$TOPIC_6$	$TOPIC_6$
	4	$TOPIC_7$	X_8	X_8
	5	$TOPIC_8$	$TOPIC_{10}$	$TOPIC_8$

According to Table 8, the Glasgow Coma Scale (X_7), Age (X_8), and Heart Rate (X_1) are relatively important variables for prediction of ICU patient mortality using structural data from ICU patients recorded within 24 h of admission. The addition of initial clinical diagnosis records (unstructured data) produced variable results about the feature significance of patient mortality prediction. In addition to the Glasgow Coma Scale (X_7), chest pain ($TOPIC_6$) and coronary artery disease ($TOPIC_1$) were also relatively significant. In the model constructed using data from ICU patients within 24 h of admission, bleed mass ($TOPIC_7$) was a relatively important variable for 3-day mortality, altered mental status ($TOPIC_{10}$) for 30-day mortality, and sepsis ($TOPIC_8$) for 365-day mortality. Important variables to consider are the Glasgow Coma Scale (X_7), chest pain ($TOPIC_6$), and coronary artery disease ($TOPIC_1$), regardless of the mortality prediction for different ICU patient time periods.

4. Discussion

4.1. Principal Findings

Previous studies have focused on building predictive models using quantitative variables from EHR databases to predict mortality, length of stay, and disease diagnosis in ICU patients. However, such studies have largely ignored the potential value of qualitative data due to challenges in standardization and utilization. By overlooking unstructured data in EHR, clinicians may miss critical information and clues provided by the physician's initial observations. To fully utilize unstructured data, this study employs NLP techniques, specifically the LDA model, to analyze clinical notes. Our study integrates structured data, such as basic vital signs, with unstructured data, derived from physicians' initial clinical diagnoses at the time of ICU admission, to predict patient mortality. Additionally, our model successfully identifies significant variables for predicting clinical outcomes during different ICU periods. We hope that our analysis results can enhance medical staff's understanding of patient conditions, optimize medical resource allocation, and provide patients, families, and medical staff with more information for informed decision-making. The main contributions of this study include: (1) investigating the impact of integrating structured and unstructured clinical records on ICU patient outcomes using a machine learning model, and (2) predicting patient mortality and risk factors to inform potential preventive measures in medical practice.

In previous studies, researchers have achieved comparable or even superior accuracy by employing excessive numbers of features. For instance, Xia et al. [13] used 50 features to achieve an AUROC of 0.85, and Liu et al. [73] employed 99 features to achieve an AUROC of 0.78. However, in our study, we achieved an accuracy of 97% and an AUROC of 0.88 for the mortality model using only six vital signs, the GCS, age, and the initial written clinical records and diagnosis made on patient admission to the ICU. We utilized eight commonly

used machine learning classification algorithms, each with a known degree of accuracy in predicting ICU patient mortality. Our AUROC findings revealed that the mortality rate in 3 days can exceed 80%, and in 30 days and 365 days can exceed 75%. Our study found that Gradient Boosting provided the most accurate prediction model. XGBoost had the highest prediction accuracy, indicating that our proposed method could predict mortality in ICU patients very well. Our results also demonstrated that the initial written notes of clinical observations and diagnoses made at the time of patient admission to the ICU contain a wealth of useful information that can aid ICU medical and nursing staff in making crucial clinical decisions. Furthermore, our study only utilized structured and unstructured data of ICU patients within 24 h of admission; our prediction model was found to be more suitable for predicting short-term mortality, as it could predict 3-day mortality with more accuracy than 30-day and 365-day mortality.

Using the LDA method, the analysis of unstructured data recorded by ICU admission clinicians during initial observation and diagnosis yielded significant results. Other important variables to consider in addition to the Glasgow Coma Scale (X_7) are patient age at admission (X_8), chest pain ($TOPIC_6$) and coronary artery disease ($TOPIC_1$). Overall, the LDA method can extract significant medical characteristics from patient topics. Furthermore, these medical characteristics can be utilized in a variety of situations to provide personalized clinical advice to individual patients [35]. In addition, various imputation techniques were applied to the dataset to determine the optimal solution for the issue at hand. Because the number of ICU patients who died in this study was significantly lower than the number of those who survived, the majority-to-minority ratio was 97 to 3 (3-day mortality) and the data demonstrate an extremely high category imbalance. SMOTE technology was used to increase the sample size of the side with the smaller number of samples to achieve data parity.

4.2. Limitations

To begin, all the data used in this paper came from a large retrospective clinical database, and the findings were generalized across groups of patients rather than specific people. To ensure the thorough collection of relevant data, this study only took into account complete patient dynamic information from databases where ICU data were easy to obtain. This study used MIMIC-III data collected at Beth Israel Deaconess Medical Center in Boston, Massachusetts. Future studies should evaluate data collected from more medical facilities across a wider geographic area. Because this study was limited to ICU medical data accumulated by a large medical facility in a big city, the findings cannot be safely applied to ICU patients in smaller medical facilities. More comprehensive results and verification could be obtained by comparing these results with those from data obtained from rural or other general small medical facilities.

Second, the model's performance may be undermined in other critical care settings due to a lack of high-quality care notes for a large number of patients. The information entered by physicians during patient consultations is valuable for disease and treatment research. Because these notes are highly telegraphic and contain many spelling errors, inconsistent punctuation, and non-standard word order, the existing natural language analysis tools struggle to process them [74]. Common spelling errors and other noise in medical notes can affect interpretation quality, resulting in counterintuitive results, which is a limitation and challenge for related research [5]. Furthermore, because this study is retrospective, conclusions about predictive algorithm performance in a hospital setting cannot be drawn. Future studies could evaluate and analyze these constraints in greater depth to make this kind of study more objective and thorough.

5. Conclusions

As the COVID-19 pandemic continues to strain ICUs worldwide, the critical importance of such facilities has become increasingly apparent. Consequently, there is a pressing need for more active research to manage scarce critical care resources and provide ad-

ditional tools to support medical decision-making and effective benchmarks for clinical practice. In this study, not only using structured data from ICU patients' first 24 h (including six vital sign measurements, the GCS, and patient age at admission), but also focusing on unstructured data from initial state observations and diagnoses made upon admission. The effectiveness of using LDA method and different machine learning technologies in the prediction of ICU patient mortality was discussed. These unstructured data contained a wealth of information that could effectively assist in later clinical decision making. However, the model developed in this study primarily focused on predicting ICU patient mortality, and further investigation is warranted to explore other clinical tasks such as length of stay, complication, and disease prediction. Moreover, it is evident that physician-produced clinical care records may capture the concepts required for mortality prediction with greater pertinence and accuracy than is currently achievable using traditional statistical techniques. Therefore, it is recommended that four directions be pursued for further research.

First, clinicians should integrate a large amount of information to evaluate and predict the current and future status of the patient, making the environment of critical care cognitively more demanding. It is essential to gain a comprehensive understanding of the specific need for clinical ICU predictive systems, the types and properties of predictions that are valued by the clinician, and the optimal time scale for such predictions. Despite the fact that findings show that our proposed method produced good predictive results for ICU patient mortality, additional research is required to evaluate its benefits on clinical care and its effectiveness to elucidate the prediction principles.

Second, the data in this study were restricted to patients admitted to the ICU for the first time and exclude patient readmission records and reports. Reduction in readmissions has long been identified by the United States government as a priority area for healthcare policy reform. Hospital readmission has also been promoted as a metric that can aid in the reduction in healthcare cost. More types of readmission research, such as the predictive performance of readmission models, could be conducted, as well as the impact of patient-level predictors on readmission, and studies of the relationship between healthcare environment quality and readmission [75,76]. Because ICU patient readmission frequently results in excessive use of medical resources and financial risk to medical facilities, analyzing the morbidity and mortality of readmitted ICU patients will benefit both patients and medical facilities [77]. Future research could collect data from multiple ICU admissions and make a comprehensive evaluation of time-series issues and also provide additional levels of analytical results as a reference for patients, medical and nursing staff, and the families of the patients.

Third, because the MIMIC-III database contains accumulated medical data from ICU patients at the medical facilities of a large city, the results of the analysis cannot be safely applied to ICU patients at smaller medical facilities. If follow-up studies are made using ICU patient data from rural and other general medical facilities as a comparison, more comprehensive results and verification would be available. Patient data should be collected from different medical centers, including outpatient, inpatient, and emergency facilities, as well as ICUs. This will allow a more comprehensive model to be constructed for evaluation and expand applicability. In addition, classification and analysis could be conducted on the basis of various diseases, such as diabetes, and disciplines such as chest medicine.

Finally, unstructured or semi-structured data account for more than 80% of the information in electronic health records. If qualitative information is ignored, clues and key factors may be missed if the initial observation-based judgments of the physician are not taken into account. In this study, unstructured data from the initial state observation and diagnosis made by physicians at ICU admission were added to the commonly used structural variables in the traditional ICU prediction model and the LDA method was used for model construction. Future research can also collect and integrate various types of unstructured data, such as the hospital consultation process, the needs of the patient, and their social media message content, to improve prediction accuracy of the model.

Other new topic modeling tools, such as BERT, can also be used to assess the power of the proposed prediction plan.

Author Contributions: Conceptualization, C.-C.C. and T.-N.C.; Data curation, C.-C.C., T.-N.C. and C.L.; Formal analysis, C.-C.C. and T.-N.C.; Methodology, C.-C.C., T.-N.C. and C.L.; Supervision, C.-C.C. and T.-N.C.; Writing—Original draft, C.-C.C. and T.-N.C.; Writing—Review and editing, C.-C.C., C.-M.W., T.-N.C., L.-J.K., C.L. and C.-M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to sincerely thank the editor and reviewers for their kind comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Confusion matrix.

		Prediction	
		Positive	**Negative**
Actual	Positive	True Positive, TP	False Negative, FN
	Negative	False Positive, FP	True Negative, TN

$$\text{Precision} = \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{A1}$$

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{A2}$$

$$\text{F1} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{A3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{A4}$$

- AUROC: The area under the ROC curve (Receiver Operating Characteristic Curve) was used to measure performance of the classifier across all classification thresholds. The AUC measuring standard assigns the same weight to each instance, regardless of the nature of the positive label. The ROC curve is obtained through the FPR value on the abscissa and the TPR value on the ordinate.
- Precision: Also called positive predictive value (PPV). This is the proportion of correct predictions in positive samples; in other words, the proportion of positive samples among all positive samples classified.
- Recall: Also called true-positive rate (TPR). This is the proportion of samples that are predicted to be correct among factual samples, or the proportion of positive samples predicted among all positive samples.
- F1-score: The harmonic mean between precision and recall. Precision and recall are often discussed in identification- and prediction-related algorithms, whereas the F1-score considers both precision and recall and is a comprehensive measure of model performance.

- Accuracy: This is the ratio of the number of samples correctly classified (by the classifier) to the total number of samples for a given test dataset. In other words, it is the overall ratio of the model that predicts the correct quantity.

Appendix B

Table A2. B-1 Diagnostic precision, recall, F1-score, and accuracy using 3-day dataset.

Dataset	Method	3 Days			
		Precision	Recall	F1-Score	Accuracy
Structured Data	AdaBoost	0.1692 ± 0.0116	0.8662 ± 0.0149	0.2828 ± 0.0156	0.8537 ± 0.0061
	Bagging	0.1370 ± 0.0062	0.9194 ± 0.0179	0.2383 ± 0.0091	0.8047 ± 0.0031
	Gradient Boosting	0.1756 ± 0.0130	0.8577 ± 0.0213	0.2911 ± 0.0175	0.8609 ± 0.0061
	LightGBM	0.2616 ± 0.0182	0.6644 ± 0.0326	0.3747 ± 0.0178	0.9263 ± 0.0029
	Logistic Regression	0.1242 ± 0.0082	0.8153 ± 0.0457	0.2156 ± 0.0140	0.8084 ± 0.0012
	MLP	0.1535 ± 0.0117	0.8525 ± 0.0281	0.2601 ± 0.0216	0.8504 ± 0.0163
	SVC	0.1244 ± 0.0074	0.8176 ± 0.0125	0.2158 ± 0.0109	0.8023 ± 0.0043
	XGBoost	0.3284 ± 0.0243	0.4240 ± 0.0256	0.3688 ± 0.0131	0.9518 ± 0.0012
Structured Data + Unstructured Data	AdaBoost	0.1752 ± 0.0015	0.8754 ± 0.0188	0.2920 ± 0.0030	0.8624 ± 0.0037
	Bagging	0.1314 ± 0.0013	0.9513 ± 0.0155	0.2309 ± 0.0023	0.8085 ± 0.0036
	Gradient Boosting	0.1863 ± 0.0060	0.8594 ± 0.0220	0.3063 ± 0.0093	0.8737 ± 0.0046
	LightGBM	0.3988 ± 0.0091	0.6975 ± 0.0299	0.5073 ± 0.0130	0.9561 ± 0.0018
	Logistic Regression	0.1264 ± 0.0013	0.8324 ± 0.0213	0.2194 ± 0.0027	0.8108 ± 0.0014
	MLP	0.1568 ± 0.0080	0.8600 ± 0.0540	0.2648 ± 0.0088	0.8543 ± 0.0181
	SVC	0.1213 ± 0.0018	0.8297 ± 0.0139	0.2117 ± 0.0032	0.8036 ± 0.0048
	XGBoost	0.5786 ± 0.0121	0.5443 ± 0.0364	0.5600 ± 0.0157	0.9723 ± 0.0005

Table A3. B-2 Diagnostic precision, recall, F1-score, and accuracy using 30-day dataset.

Dataset	Method	30 Days			
		Precision	Recall	F1-Score	Accuracy
Structured Data	AdaBoost	0.2872 ± 0.0116	0.7252 ± 0.0181	0.4114 ± 0.0148	0.7719 ± 0.0029
	Bagging	0.2880 ± 0.0097	0.7709 ± 0.0092	0.4193 ± 0.0116	0.7653 ± 0.0034
	Gradient Boosting	0.2904 ± 0.0152	0.7262 ± 0.0213	0.4149 ± 0.0190	0.7748 ± 0.0059
	LightGBM	0.3349 ± 0.0081	0.6875 ± 0.0148	0.4504 ± 0.0104	0.8156 ± 0.0014
	Logistic Regression	0.2717 ± 0.0081	0.7462 ± 0.0165	0.3983 ± 0.0109	0.7522 ± 0.0006
	MLP	0.2846 ± 0.0329	0.7520 ± 0.0265	0.4113 ± 0.0310	0.7610 ± 0.0318
	SVC	0.2854 ± 0.0107	0.7347 ± 0.0205	0.4111 ± 0.0143	0.7686 ± 0.0027
	XGBoost	0.3753 ± 0.0129	0.5577 ± 0.0157	0.4486 ± 0.0128	0.8493 ± 0.0024
Structured Data + Unstructured Data	AdaBoost	0.2865 ± 0.0026	0.7275 ± 0.0145	0.4111 ± 0.0050	0.7740 ± 0.0012
	Bagging	0.2873 ± 0.0013	0.7531 ± 0.0264	0.4158 ± 0.0049	0.7696 ± 0.0032
	Gradient Boosting	0.2928 ± 0.0007	0.7335 ± 0.0190	0.4185 ± 0.0033	0.7780 ± 0.0021
	LightGBM	0.3411 ± 0.0014	0.6754 ± 0.0289	0.4532 ± 0.0078	0.8226 ± 0.0021
	Logistic Regression	0.2664 ± 0.0046	0.7291 ± 0.0205	0.3902 ± 0.0077	0.7568 ± 0.0039
	MLP	0.2689 ± 0.0097	0.7549 ± 0.0492	0.3958 ± 0.0059	0.7686 ± 0.0210
	SVC	0.2685 ± 0.0046	0.7250 ± 0.0236	0.3918 ± 0.0083	0.7649 ± 0.0011
	XGBoost	0.3816 ± 0.0061	0.5435 ± 0.0137	0.4483 ± 0.0082	0.8543 ± 0.0015

Table A4. B-3 Diagnostic precision, recall, F1-score, and accuracy using 365-day dataset.

Dataset	Method	365 Days			
		Precision	Recall	F1-Score	Accuracy
Structured Data	AdaBoost	0.3005 ± 0.0030	0.7125 ± 0.0030	0.4227 ± 0.0028	0.7750 ± 0.0050
	Bagging	0.2919 ± 0.0022	0.7397 ± 0.0117	0.4186 ± 0.0041	0.7625 ± 0.0029
	Gradient Boosting	0.3043 ± 0.0021	0.7097 ± 0.0014	0.4259 ± 0.0023	0.7789 ± 0.0013

Table A4. *Cont.*

Dataset	Method	365 Days			
		Precision	Recall	F1-Score	Accuracy
Structured Data	LightGBM	0.3438 ± 0.0057	0.6723 ± 0.0073	0.4549 ± 0.0063	0.8138 ± 0.0028
	Logistic Regression	0.2763 ± 0.0010	0.7156 ± 0.0096	0.3986 ± 0.0017	0.7504 ± 0.0038
	MLP	0.2918 ± 0.0141	0.7325 ± 0.0389	0.4166 ± 0.0100	0.7622 ± 0.0213
	SVC	0.2917 ± 0.0013	0.7159 ± 0.0089	0.4145 ± 0.0023	0.7662 ± 0.0044
	XGBoost	0.3875 ± 0.0035	0.5571 ± 0.0108	0.4570 ± 0.0055	0.8470 ± 0.0017
Structured Data + Unstructured Data	AdaBoost	0.3018 ± 0.0098	0.7462 ± 0.0197	0.4298 ± 0.0132	0.7758 ± 0.0053
	Bagging	0.3054 ± 0.0057	0.7681 ± 0.0118	0.4370 ± 0.0066	0.7760 ± 0.0024
	Gradient Boosting	0.3065 ± 0.0083	0.7463 ± 0.0141	0.4345 ± 0.0107	0.7801 ± 0.0055
	LightGBM	0.3602 ± 0.0056	0.6997 ± 0.0068	0.4755 ± 0.0061	0.8253 ± 0.0012
	Logistic Regression	0.2822 ± 0.0055	0.7381 ± 0.0094	0.4083 ± 0.0071	0.7578 ± 0.0020
	MLP	0.2826 ± 0.0039	0.7518 ± 0.0165	0.4107 ± 0.0046	0.7657 ± 0.0079
	SVC	0.2866 ± 0.0058	0.7306 ± 0.0055	0.4117 ± 0.0068	0.7696 ± 0.0042
	XGBoost	0.3945 ± 0.0033	0.5659 ± 0.0112	0.4649 ± 0.0060	0.8526 ± 0.0017

References

1. Marshall, J.C.; Bosco, L.; Adhikari, N.K.; Connolly, B.; Diaz, J.V.; Dorman, T.; Fowler, R.A.; Meyfroidt, G.; Nakagawa, S.; Pelosi, P.; et al. What is an intensive care unit? A report of the task force of the World Federation of Societies of Intensive and Critical Care Medicine. *J. Crit. Care* **2017**, *37*, 270–276. [CrossRef] [PubMed]
2. Mahbub, M.; Srinivasan, S.; Danciu, I.; Peluso, A.; Begoli, E.; Tamang, S.; Peterson, G.D. Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PLoS ONE* **2022**, *17*, e0262182.
3. Chen, W.; Long, G.; Yao, L.; Sheng, Q.Z. AMRNN: Attended multi-task recurrent neural networks for dynamic illness severity prediction. *World Wide Web* **2019**, *23*, 2753–2770. [CrossRef]
4. Romana, S.; Bernhard, F. Iatrogenic events contributing to paediatric intensive care unit admission. *Swiss Med. Wkly.* **2021**, *151*, 7.
5. Caicedo-Torres, W.; Gutierrez, K. ISeeU2: Visually interpretable mortality prediction inside the ICU using deep learning and free-text medical notes. *Expert Syst. Appl.* **2022**, *202*, 117190. [CrossRef]
6. Romano, M. The Role of Palliative Care in the Cardiac Intensive Care Unit. *Healthcare* **2019**, *7*, 30. [CrossRef]
7. El-Rashidy, N.; El-Sappagh, S.; Abuhmed, T.; Abdelrazek, S.; El-Bakry, H.M. Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model. *IEEE Access* **2020**, *8*, 133541–133564. [CrossRef]
8. Vincent, J.L.; Moreno, R.; Takala, J.; Willatts, S.; De Mendonça, A.; Bruining, H.; Reinhart, C.K.; Suter, P.; Thijs, L.G. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* **1996**, *22*, 707–710. [CrossRef]
9. Legall, J.R.; Lemeshow, S.; Saulnier, F. A new simplified acute physiology score (SAPS-II) based on a European North-American multicenter study. *Jama J. Am. Med. Assoc.* **1993**, *270*, 2957–2963. [CrossRef]
10. Baue, A.E.; Durham, R.; Faist, E. Systemic inflammatory response syndrome (SIRS), multiple organ dysfunction syndrome (MODS), multiple organ failure (MOF): Are we winning the battle? *Shock* **1998**, *10*, 79–89. [CrossRef]
11. Ibrahim, Z.M.; Wu, H.H.; Hamoud, A.; Stappen, L.; Dobson, R.J.B.; Agarossi, A. On classifying sepsis heterogeneity in the ICU: Insight using machine learning. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 437–443. [CrossRef] [PubMed]
12. Darabi, S.; Kachuee, M.; Fazeli, S.; Sarrafzadeh, M. TAPER: Time-Aware Patient EHR Representation. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3268–3275. [CrossRef] [PubMed]
13. Gong, M.G.; Pan, K.; Xie, Y.; Qin, A.K.; Tang, Z.D. Preserving differential privacy in deep neural networks with relevance-based adaptive noise imposition. *Neural Netw.* **2020**, *125*, 131–141. [CrossRef] [PubMed]
14. Sheikhalishahi, S.; Balaraman, V.; Osmani, V. Benchmarking machine learning models on multi-centre eICU critical care dataset. *PLoS ONE* **2020**, *15*, e0235424. [CrossRef]
15. Loreto, M.; Lisboa, T.; Moreira, V.P. Early prediction of ICU readmissions using classification algorithms. *Comput. Biol. Med.* **2020**, *118*, 8. [CrossRef]
16. Baker, S.; Xiang, W.; Atkinson, I. Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: A hybrid neural network approach. *Sci. Rep.* **2020**, *10*, 1–12. [CrossRef]
17. Davidson, S.; Villarroel, M.; Harford, M.; Finnegan, E.; Jorge, J.; Young, D.; Watkinson, P.; Tarassenko, L. Day-to-day progression of vital-sign circadian rhythms in the intensive care unit. *Crit. Care* **2021**, *25*, 13. [CrossRef]
18. Alghatani, K.; Ammar, N.; Rezgui, A.; Shaban-Nejad, A. Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation. *JMIR Med. Inform.* **2021**, *9*, e21347. [CrossRef]

19. Sarang, B.; Bhandarkar, P.; Raykar, N.; O'Reilly, G.M.; Soni, K.D.; Wärnberg, M.G.; Khajanchi, M.; Dharap, S.; Cameron, P.; Howard, T.; et al. Associations of On-arrival Vital Signs with 24-hour In-hospital Mortality in Adult Trauma Patients Admitted to Four Public University Hospitals in Urban India: A Prospective Multi-Centre Cohort Study. *Inj. Int. J. Care Inj.* **2021**, *52*, 1158–1163. [CrossRef]
20. Hashir, M.; Sawhney, R. Towards unstructured mortality prediction with free-text clinical notes. *J. Biomed. Inform.* **2020**, *108*, 103489. [CrossRef]
21. Tootooni, M.S.; Pasupathy, K.S.; Heaton, H.A.; Clements, C.M.; Sir, M.Y. CCMapper: An adaptive NLP-based free-text chief complaint mapping algorithm. *Comput. Biol. Med.* **2019**, *113*, 13. [CrossRef] [PubMed]
22. Ye, J.C.; Yao, L.; Shen, J.H.; Janarthanam, R.; Luo, Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 7. [CrossRef] [PubMed]
23. Zhang, D.D.; Yin, C.C.; Zeng, J.C.; Yuan, X.H.; Zhang, P. Combining structured and unstructured data for predictive models: A deep learning approach. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 280. [CrossRef] [PubMed]
24. Mitchell, T. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997; Volume 1.
25. Adlung, L.; Cohen, Y.; Mor, U.; Elinav, E. Machine learning in clinical decision making. *Med* **2021**, *2*, 642–665. [CrossRef]
26. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef]
27. Purushotham, S.; Meng, C.Z.; Che, Z.P.; Liu, Y. Benchmarking deep learning models on large healthcare datasets. *J. Biomed. Inform.* **2018**, *83*, 112–134. [CrossRef]
28. Cheng, X.; Cao, Q.; Liao, S.S. An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. *J. Inf. Sci.* **2022**, *48*, 304–320. [CrossRef]
29. Xue, J.; Chen, J.X.; Chen, C.; Zheng, C.D.; Li, S.J.; Zhu, T.S. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PLoS ONE* **2020**, *15*, e0239441. [CrossRef]
30. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
31. Breuninger, T.A.; Wawro, N.; Breuninger, J.; Reitmeier, S.; Clavel, T.; Six-Merker, J.; Pestoni, G.; Rohrmann, S.; Rathmann, W.; Peters, A.; et al. Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. *Microbiome* **2021**, *9*, 61. [CrossRef]
32. Gangavarapu, T.; Jayasimha, A.; Krishnan, G.S.; Kamath, S.S. Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowl. Based Syst.* **2020**, *190*, 105321. [CrossRef]
33. Chiu, C.C.; Wu, C.M.; Chien, T.N.; Kao, L.J.; Qiu, J.T. Predicting the Mortality of ICU Patients by Topic Model with Machine-Learning Techniques. *Healthcare* **2022**, *10*, 1087. [CrossRef] [PubMed]
34. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.W.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
35. Yu, R.; Zheng, Y.; Zhang, R.; Jiang, Y.; Poon, C.C.Y. Using a Multi-Task Recurrent Neural Network With Attention Mechanisms to Predict Hospital Mortality of Patients. *IEEE J. Biomed. Health Inf.* **2020**, *24*, 486–492. [CrossRef]
36. Guo, C.H.; Lu, M.L.; Chen, J.F. An evaluation of time series summary statistics as features for clinical prediction tasks. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 48. [CrossRef]
37. Sayed, M.; Riano, D.; Villar, J. Predicting Duration of Mechanical Ventilation in Acute Respiratory Distress Syndrome Using Supervised Machine Learning. *J. Clin. Med.* **2021**, *10*, 3824. [CrossRef]
38. Kozlowski, D.; Semeshenko, V.; Molinari, A. Latent Dirichlet allocation model for world trade analysis. *PLoS ONE* **2021**, *16*, e0245393. [CrossRef]
39. Li, Y.; Rapkin, B.; Atkinson, T.M.; Schofield, E.; Bochner, B.H. Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. *Qual. Life Res.* **2019**, *28*, 1441–1455. [CrossRef]
40. Celard, P.; Vieira, A.S.; Iglesias, E.L.; Borrajo, L. LDA filter: A Latent Dirichlet Allocation preprocess method for Weka. *PLoS ONE* **2020**, *15*, e0241701. [CrossRef]
41. Chen, J.H.; Goldstein, M.K.; Asch, S.M.; Mackey, L.; Altman, R.B. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 472–480. [CrossRef]
42. Pivovarov, R.; Perotte, A.J.; Grave, E.; Angiolillo, J.; Wiggins, C.H.; Elhadad, N. Learning probabilistic phenotypes from heterogeneous EHR data. *J. Biomed. Inform.* **2015**, *58*, 156–165. [CrossRef]
43. Choi, Y.; Chiu, C.Y.-I.; Sontag, D. Learning low-dimensional representations of medical concepts. *AMIA Summits Transl. Sci. Proc.* **2016**, *2016*, 41–50. [PubMed]
44. Gabriel, R.A.; Kuo, T.-T.; McAuley, J.; Hsu, C.-N. Identifying and characterizing highly similar notes in big clinical note datasets. *J. Biomed. Inform.* **2018**, *82*, 63–69. [CrossRef] [PubMed]
45. Teng, F.; Ma, Z.; Chen, J.; Xiao, M.; Huang, L.F. Automatic Medical Code Assignment via Deep Learning Approach for Intelligent Healthcare. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2506–2515. [CrossRef] [PubMed]
46. Kim, D.H.; Choi, J.Y.; Ro, Y.M. Region based stellate features combined with variable selection using AdaBoost learning in mammographic computer-aided detection. *Comput. Biol. Med.* **2015**, *63*, 238–250. [CrossRef] [PubMed]
47. Lee, Y.W.; Choi, J.W.; Shin, E.H. Machine learning model for predicting malaria using clinical information. *Comput. Biol. Med.* **2021**, *129*, 104151. [CrossRef]
48. Ali, S.; Majid, A.; Javed, S.G.; Sattar, M. Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data. *Comput. Biol. Med.* **2016**, *73*, 38–46. [CrossRef]

49. Sarmah, C.K.; Samarasinghe, S. Microarray gene expression: A study of between-platform association of Affymetrix and cDNA arrays. *Comput. Biol. Med.* **2011**, *41*, 980–986. [CrossRef]
50. Ramos-Gonzalez, J.; Lopez-Sanchez, D.; Castellanos-Garzon, J.A.; de Paz, J.F.; Corchado, J.M. A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. *Comput. Biol. Med.* **2017**, *86*, 98–106. [CrossRef]
51. Song, J.Z.; Liu, G.X.; Jiang, J.Q.; Zhang, P.; Liang, Y.C. Prediction of Protein-ATP Binding Residues Based on Ensemble of Deep Convolutional Neural Networks and LightGBM Algorithm. *Int. J. Mol. Sci.* **2021**, *22*, 939. [CrossRef]
52. Li, L.J.; Lin, Y.K.; Yu, D.X.; Liu, Z.Y.; Gao, Y.J.; Qiao, J.P. A Multi-Organ Fusion and LightGBM Based Radiomics Algorithm for High-Risk Esophageal Varices Prediction in Cirrhotic Patients. *IEEE Access* **2021**, *9*, 15041–15052. [CrossRef]
53. Cuadrado-Godia, E.; Jamthikar, A.D.; Gupta, D.; Khanna, N.N.; Araki, T.; Maniruzzaman, M.; Saba, L.; Nicolaides, A.; Sharma, A.; Omerzu, T.; et al. Ranking of stroke and cardiovascular risk factors for an optimal risk calculator design: Logistic regression approach. *Comput. Biol. Med.* **2019**, *108*, 182–195. [CrossRef] [PubMed]
54. Ergun, U.; Serhatioglu, S.; Hardalac, F.; Guler, I. Classification of carotid artery stenosis of patients with diabetes by neural network and logistic regression. *Comput. Biol. Med.* **2004**, *34*, 389–405. [CrossRef] [PubMed]
55. Kavitha, M.S.; Kurita, T.; Ahn, B.C. Critical texture pattern feature assessment for characterizing colonies of induced pluripotent stem cells through machine learning techniques. *Comput. Biol. Med.* **2018**, *94*, 55–64. [CrossRef] [PubMed]
56. Guler, E.C.; Sankur, B.; Kahya, Y.P.; Raudys, S. Visual classification of medical data using MLP mapping. *Comput. Biol. Med.* **1998**, *28*, 275–287. [CrossRef] [PubMed]
57. Nanayakkara, S.; Fogarty, S.; Tremeer, M.; Ross, K.; Richards, B.; Bergmeir, C.; Xu, S.; Stub, D.; Smith, K.; Tacey, M.; et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS Med.* **2018**, *15*, e1002709. [CrossRef]
58. Akbari, G.; Nikkhoo, M.; Wang, L.; Chen, C.P.; Han, D.S.; Lin, Y.H.; Chen, H.B.; Cheng, C.H. Frailty Level Classification of the Community Elderly Using Microsoft Kinect-Based Skeleton Pose: A Machine Learning Approach. *Sensors* **2021**, *21*, 4017. [CrossRef]
59. Hou, N.; Li, M.; He, L.; Xie, B.; Wang, L.; Zhang, R.; Yu, Y.; Sun, X.; Pan, Z.; Wang, K. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost. *J. Transl. Med.* **2020**, *18*, 462. [CrossRef]
60. Luo, X.Q.; Yan, P.; Duan, S.B.; Kang, Y.X.; Deng, Y.H.; Liu, Q.; Wu, T.; Wu, X. Development and Validation of Machine Learning Models for Real-Time Mortality Prediction in Critically Ill Patients With Sepsis-Associated Acute Kidney Injury. *Front. Med.* **2022**, *9*, 853102. [CrossRef]
61. Raghuwanshi, B.S.; Shukla, S. Classifying imbalanced data using SMOTE based class-specific kernelized ELM. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1255–1280. [CrossRef]
62. Zhang, Y.; Jiang, Z.W.; Chen, C.; Wei, Q.Q.; Gu, H.M.; Yu, B. DeepStack-DTIs: Predicting Drug-Target Interactions Using LightGBM Feature Selection and Deep-Stacked Ensemble Classifier. *Interdiscip. Sci. Comput. Life Sci.* **2022**, *14*, 311–330. [CrossRef]
63. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
64. Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Machine learning and statistical methods for predicting mortality in heart failure. *Heart Fail. Rev.* **2021**, *26*, 545–552. [CrossRef] [PubMed]
65. Javan, S.L.; Sepehri, M.M.; Javan, M.L.; Khatibi, T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. *Comput. Methods Programs Biomed.* **2019**, *178*, 47–58. [CrossRef] [PubMed]
66. Blagus, R.; Lusa, L. Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models. *BMC Bioinform.* **2015**, *16*, 363. [CrossRef] [PubMed]
67. Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chen, J.; Chou, K.-C. Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS ONE* **2015**, *10*, e0121501. [CrossRef]
68. Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chou, K.-C. iMiRNA-PseDPC: MicroRNA precursor identification with a pseudo distance-pair composition approach. *J. Biomol. Struct. Dyn.* **2016**, *34*, 223–235. [CrossRef]
69. Upadhyay, D.; Manero, J.; Zaman, M.; Sampalli, S. Gradient Boosting Feature Selection With Machine Learning Classifiers for Intrusion Detection on Power Grids. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 1104–1116. [CrossRef]
70. Adler, A.I.; Painsky, A. Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy* **2022**, *24*, 687. [CrossRef]
71. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2001.
72. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
73. Liu, D.; Wu, Y.L.; Li, X.; Qi, L. Medi-Care AI: Predicting medications from billing codes via robust recurrent neural networks. *Neural Netw.* **2020**, *124*, 109–116. [CrossRef]
74. Savkov, A.; Carroll, J.; Koeling, R.; Cassell, J. Annotating patient clinical records with syntactic chunks and named entities: The Harvey Corpus. *Lang. Resour. Eval.* **2016**, *50*, 523–548. [CrossRef] [PubMed]
75. Qiu, L.F.; Kumar, S.; Sen, A.; Sinha, A. Impact of the Hospital Readmission Reduction Program on hospital readmission and mortality: An economic analysis. *Prod. Oper. Manag.* **2022**, *31*, 2341–2360. [CrossRef]

76. Senot, C. Continuity of care and risk of readmission: An investigation into the healthcare journey of heart failure patients. *Prod. Oper. Manag.* **2019**, *28*, 2008–2030. [CrossRef]
77. Lin, Y.W.; Zhou, Y.Q.; Faghri, F.; Shawl, M.J.; Campbell, R.H. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long shortterm memory. *PLoS ONE* **2019**, *14*, e0218942.