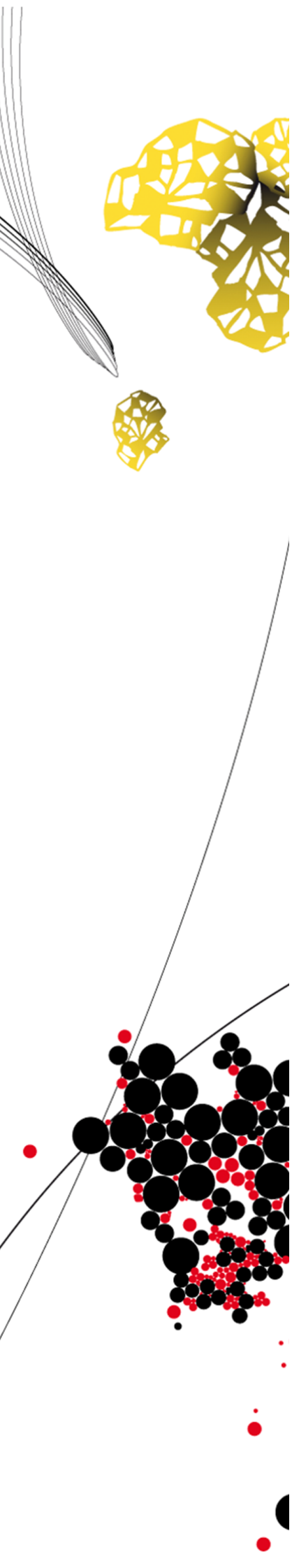

UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Incorporating Clinical Notes in an Early Sepsis Prediction Model to Improve Performance

Siba Siddique

M.Sc. Thesis in Interaction Technology

September 2020

Supervisors:
dr. F. Shamout (New York University Abu Dhabi)
dr. M. Theune
dr. S. Wang

Human Media Interaction Group
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Acknowledgements

This research internship would not have been possible without the collaboration between New York University Abu Dhabi (NYUAD) and Cleveland Clinic Abu Dhabi (CCAD).

I would firstly like to thank my supervisors, Dr. Farah Shamout from the Clinical Artificial Intelligence Research Laboratory at NYUAD, Dr. Shenghui Wang and Dr. Mariët Theune from the Human Media Interaction Group at the University of Twente, for their unwavering support and clear feedback that pushed me to do my best. I feel greatly privileged to have learned from the best mentors in the fields of AI and NLP, who provided me with the tools for success despite the challenges posed by the Covid-19 pandemic.

Thank you to all my mentors and fellow classmates at EIT Digital, University Paris-Sud, and University of Twente. It has been an incredible learning experience thus far, and I am excited for what is to come!

I thank the Almighty for His countless blessings and for giving me the determination and strength to do my research. Last but not least, I extend my heartfelt gratitude to my family, for supporting me in every step of my journey to my Master's degree.

Abstract

Sepsis is a condition caused by the body's response to an infection that affects an estimated 50 million people globally and is one of the leading contributors to hospital mortality. Risk-prediction models for sepsis onset prediction are currently used in hospitals to assist with clinical decision-making. The majority of these systems have been developed using machine learning and gradient-boosting algorithms on electronic health records (EHR) data. Recently, the use of natural language processing (NLP) on clinical notes has aided in improving patient outcome predictions and identifying patient diagnosis.

The aim of this study is to improve the prediction of sepsis onset by combining clinical notes as an added modality with the structured data components of the EHR data. Our model is trained on the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, and explores the use of time-series physiological data with clinical note embeddings. To assess the effect of the input features we evaluated logistic regression, multinomial Naïve Bayes, and XGBoost (XGB) models on the following three configurations: (1) structured EHR data (physiological measures), (2) clinical note embeddings alone, and (3) the combination of physiological and note features. Furthermore, we assessed the effect of using different prediction time and look back intervals of time-series physiological data with clinical note embeddings. Pointwise mutual information (PMI) was used to find the top 200 informative words relating to sepsis from all the unique notes per admission. We compared three methods of clinical notes representations: (1) bag-of-words (BOW) model that included the top 200 PMI vectors, (2) term-frequency inverse-document frequency (Tf-idf) weighted PMI vectors, and (3) pre-trained note embeddings of 200 dimensions.

The best-performing model was the XGB model trained on the combined physiological features and pre-trained note embeddings for a 24 hour look back and prediction time interval. Finally, we propose methods to improve the acceptability and implementation of such systems in a hospital ICU setting using the findings from this research.

***K*eywords** artificial intelligence · sepsis · risk prediction · clinical decision support systems · electronic health records

Contents

List of acronyms

BSPM	benchmark sepsis prediction model
CDS	clinical decision support
EHR	electronic health records
EMR	electronic medical records
HIS	hospital information system
ICD-9	International Classification of Diseases, Ninth Revision
ICU	intensive care unit
LOS	length of stay
SAPS-II	Simplified Acute Physiology Score
SIRS	systemic inflammatory response syndrome
SOFA	sequential organ failure assessment
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
WBC	white blood cell
iDASH	Integrating Data for Analysis, Anonymization, and Sharing
MGH	Massachusetts General Hospital
MIMIC-III	Medical Information Mart for Intensive Care III
NLP	natural language processing
ROC	receiver operating characteristic
PCA	principal component analysis

LSA latent semantic analysis

LDA linear discriminant analysis

PMI pointwise mutual information

SVD singular value decomposition

Tf-idf term frequency-inverse document frequency

ML machine learning

AUROC Area under the Receiver Operating Characteristic

AUPRC Area under the Precision-Recall Curve

BERT Bidirectional Encoder Representations from Transformers

BOW bag-of-words

cTAKES clinical Text Analysis and Knowledge Extraction System

GBM Gradient Boosting Machine

GloVe Global Vectors for Word Representation

MGP Multitask Gaussian Process

LSTM long short-term memory

CNN convolutional neural network

RNN recurrent neural network

SHAP SHapley Additive exPlanations

SVM support vector machine

ROC receiver operating characteristic

UMLS Unified Medical Language System

Chapter 1

Introduction

1.1 Motivation

Sepsis is a severe bacterial infection in the blood that affects 50 million people globally each year, and accounted for 20% of deaths globally in 2017 (42). A delay in treatment could lead to serious morbidity and mortality, as each hour of delayed treatment for septic patients increased patient mortality by 8% compared to the survival rate of 80% for a patient treated within the first hour of diagnosis (24). Therefore, it is necessary to be able to detect symptoms ahead of time for a given patient population (14).

Based on our discussions with clinical domain experts at Cleveland Clinic Abu Dhabi in the United Arab Emirates (UAE), there is a keen interest in developing an early detection of sepsis model to improve patient outcomes due to the high prevalence of sepsis among acute care patients in the UAE (14 – 17% compared to 2 – 4% in the United States). Moreover, the disease occurs mostly among people who have lower than normal levels of immunity, or high-risk groups.

We used a proprietary algorithm from a leading hospital information system (HIS) vendor that predicts sepsis four hours in advance based on commonly available structured EHR data including laboratory values, comorbidity, and procedural variables. The details of the benchmark sepsis prediction model (BSPM) are not published because it is a proprietary model. Although past studies have addressed the effect of having different prediction and look back time intervals, they have not combined it with clinical note embeddings (45). Furthermore, using time-series analysis with evidence-based practice has shown to "provide approximately 20% improvement over traditional indices of heart rate entropy in the Area under the Receiver Operating Characteristic (AUROC) for four-hour advance prediction of sepsis" (47).

Moreover, it has been found that the combination of natural language processing (NLP) and physiological features in prediction tasks (such as predicting in-hospital mortality) achieved greater performance compared to using either the physiological

features or NLP features alone (30). Therefore, we are interested in investigating whether useful information pertaining to a patient's sepsis risk score can be obtained from doctor's notes; consequently, providing doctors with a better estimate of a patient's score.

1.2 Research Questions

The aim of this thesis research is to improve the performance of an early sepsis prediction model through incorporating clinical notes. We hypothesize that learning from both clinical notes and physiological variables (i.e. vital signs, laboratory results, and medications) will improve the performance of early sepsis prediction through the use of NLP and other machine learning (ML) techniques. To this end, we develop a novel sepsis prediction model that learns from physiological factors and free-text clinical notes from the MIMIC-III dataset (20).

We aim to address the following research questions in our study:

- How can we improve the existing BSPM performance by incorporating clinical notes with time-series physiological measures on an unseen dataset?
  - Which clinical note representations and shallow ML models are best suited to predict sepsis onset in adult ICU patients?
  - What effect do longer look back and prediction windows have on the model performance?

1.3 Research Tasks

In order to address the research questions, we looked at three different ML models: logistic regression, multinomial Naïve Bayes (MNB), and tree-based such as XGBoost (5). Firstly, we investigated the effect of using time series data compared to single time instant data on the physiological measures within different prediction times and different look back intervals. Secondly, three novel document encoding approaches for the clinical notes were juxtaposed to see which one gives the best prediction for the onset of sepsis. Thirdly, we compared the performances of running the models trained on the different feature sets: (a) structured data components of the EHR (physiological measures), (b) clinical note embeddings alone, and (c) multimodal system with a combination of structured EHR and note features.

The findings support our hypothesis that incorporating structured EHR data with clinical note embeddings improved the performance of the model, with the highest performance achieved using pre-trained word embeddings, followed by the Tf-idf

weighted one-hot encoded vectors. The XGB model achieved the highest AUROC and AUPRC scores overall on the combined physiological features with the pre-trained note embeddings for the 24 hour prediction time interval. Moreover, the longer history (look back) of the time series variables were also shown to improve the model performance.

1.4 Thesis Structure

The remainder of this thesis is organized as follows. In Chapter 2, we review the state-of-the-art works related to NLP applications and early prediction systems. In Chapter 3, we propose three document embedding approaches and describe the data preprocessing pipeline for the physiological factors and notes. The experiment results and discussion are summarized in Chapters 4 and 5, respectively. Finally, Chapter 6 outlines the conclusions and recommendations for future work.

Chapter 2

Related Work

This chapter presents the existing literature on the topic of machine learning prediction models within the clinical domain, specifically for sepsis prediction. Section 2.1 outlines the standard scores that are used to discern whether a patient has sepsis, such as systemic inflammatory response syndrome (SIRS) and sequential organ failure assessment (SOFA). Section 2.2 presents the machine learning models for sepsis prediction. Section 2.3 introduces the advantages and the various applications of using NLP within healthcare, and section 2.4 presents the use of word embedding models in different prediction tasks.

2.1 Early Prediction Models for Sepsis

Early detection of sepsis can help prevent more serious complications from arising. However, the main challenge is that it is hard to predict sepsis occurrence with certainty (13). Recently, many algorithms and machine learning (including deep learning) models were proposed to help predict the onset of sepsis (22, 45). A systematic review of studies targeting sepsis in a hospital setting showed that temperature, lab values, and model type were the main contributors to model performance (15).

The SOFA is an integrated score that helps determine the extent of rate of organ failure by tracking a person's status and is used for intensive care unit (ICU) mortality prediction. It is also used for sepsis prediction which is indicated by a change of two or more in the SOFA score (25). A score of zero (normal function) to four (abnormal function) is assigned to each of six organ systems (respiratory, coagulation, hepatic, cardiovascular, central nervous system and renal), resulting in a final score ranging from 0-24.

Over the years, there have been many revisions to the sepsis definition. In 1992, Bone et al. (3) identified sepsis in a patient if they had suspected infection and satisfied two out of the four SIRS criteria shown in Table 2.1. Sepsis-3 is the most recent

definition of sepsis developed in 2016 that identified patients at risk due to sepsis if there is an increase in the SOFA score of two points or more, and probable or confirmed infection. The qSOFA (or quick SOFA) is a simplified score used for sepsis prediction, with the score ranging from 0-3. A score of two or more is associated with a greater risk of death or poor outcome (46).

Table 2.1: The diagnosis for SIRS is established when there are two or more co-existing conditions, as shown in the table. Sepsis occurs when SIRS is induced as a result of infection

Factors	**Conditions**
Temperature (°C)	< 36°C or > 38°C
Heart Rate	>90/min
Respiratory rate or PCO_2	>20/min <32 mmHg
White blood cell count or immature band forms	<4k/uL or >12k/uL >10%

The limitation of the Sepsis-3 criteria is that it was not intended to be used as a clinical decision support (CDS) tool in the ICU due to "requiring the presence of organ failure" which could "delay treatment of patients who might benefit from an early approach" (9). Furthermore, the previous sepsis definition based on the presence of the SIRS continues to be used by the Centers for Medicare and Medicaid Services (CMS) to measure compliance with the sepsis quality of care bundles until it is shown that the newer definition is superior in predicting the onset of sepsis in patients (9).

2.2 Machine Learning Models for Sepsis Prediction

The MIMIC-III dataset has been used to develop baseline methods for a variety of tasks: prediction of mortality from early admission data (classification), real-time detection of decompensation (time series classification), forecasting length of stay (regression) and phenotype classification (multilabel sequence classification)(16). Machine learning models to predict the onset of sepsis are usually left or right aligned, which refer to making the prediction at the time of admission or after a given period of time, respectively (15). Right-aligned models are also known as real-time or continuous prediction models and will be the focus of our study.

InSight is a ML-based sepsis prediction algorithm that uses nine commonly available vital signs to compute a real-time risk score, and predicts sepsis onset at least three hours prior to a sustained SIRS event (4). The study was done on adult patients from the MIMIC-II v3 database who were not septic at the time of admission. The observations were rounded to the nearest hour and modeled as a causal time-series data (4). The InSight scores were calculated using a higher-order equation, which worked better than lower order trends which had higher sensitivity and led to increased false positive rates (4).

Mao et al. (2017) conducted a multicentre validation of InSight with the gradient tree boosting, or XGBoost(5), model using the time series data of six vital signs to predict and detect sepsis, severe sepsis and septic shock. Their experiments also explored the effectiveness of the model in cases where sepsis prevalence is lower (less effective) or higher (more effective) compared to the dataset they trained on. The model achieves an AUROC curve of 0.92 (95% CI 0.90 to 0.93) for sepsis detection (29). Scherpf et al. (2019) implemented a sepsis prediction tool using the MIMIC-III Database, and compared its performance with InSight (20, 45). Their study investigated the effect of different prediction times (3, 6, 12 hours) and the length of the look back (5, 10, 15, 20 hours) (45). Their model showed the importance of having a longer look back due to the ability to exploit time-dependent patterns from the “symptoms and related vital sign patterns of sepsis" which are detected by ML algorithms (45). However, the model’s black box nature restricts its usage as an early warning system rather than a decision system, and limits its interpretability compared to other interpretable models.

A Multitask Gaussian Process (MGP) recurrent neural network (RNN) classifier for the prediction of sepsis onset was developed by Bedoya et al. (2020) using 86 variables including patient demographics, comorbidities, vital signs, medications, and labs from an academic hospital (1). While RNN typically require evenly spaced inputs, using MGP with the RNN handles the irregular spacing and missing values in the raw data by maintaining the uncertainty about the variance of the series at each point (1). This is important when working with sparse data as there is a better imputation of continuous functions for all vital signs and laboratory measurements. The study also showed that "SIRS consistently outperforms qSOFA in detecting sepsis early" (1).

Shashikumar et al.’s team captured the physiological state trajectory through time using time-lagged embedding and a multiscale network representation (MSNR) (47). The patient data was extracted from an Emory affiliated ICU, and each patient’s heart rate (HR) and mean arterial blood pressure (MAP) time series at 2 second resolution were rank order transformed to eight levels, and time-synchronized with the patient’s electronic medical records (EMR) data. This was followed by partition-

ing the state-space into time-varying bins which are transformed to a network from which different topological attributes can be derived and used as input to train a support vector machine (SVM) classification model. The performances of the classifier trained on different combinations of network, multiscale entropy, and EMR features were compared. The model with the combined features achieved the highest AUROC (47).

Yu et al. (2020) developed a framework for dynamic monitoring of ICU patients' mortality risk that used the BOW representation with a long short-term memory (LSTM) RNN on the MIMIC-III dataset (53). The model is robust to missing data and uses latent semantic analysis (LSA) to encode the patient's state by taking the BOW representations and applying singular value decomposition (SVD) to perform dimensionality reduction and simplify matrix calculations (53). They consider a fixed history and prediction window of 48 and 12 hours, respectively. Different architectures were explored using a logistic regression model for binary classification, and the bi-directional LSTM model achieved the highest performance compared to the existing severity scoring system, Simplified Acute Physiology Score (SAPS-II).

The sepsis prediction model that we consider as a benchmark (BSPM) learns from the following clinical measurements: demographics, vital signs, laboratory test results and medication orders. A patient is determined to have sepsis through a list of diagnosis codes four hours before the first clinical intervention is taken or documented. The list of interventions may be one of the following: positive documentation of sepsis or suspicion of sepsis, or an order for a lactate lab or one of a few specific antibiotics used in treatment of sepsis.

2.3 Natural Language Processing applications

The use of NLP within healthcare applications can greatly reduce the time required to analyse large collections of textual data, by extracting the meaning of the text for downstream classification tasks. For example, it can notify providers about the prevalence of a specific disease through topic modelling by capturing key symptoms of the disease from clinical text, generate "domain-aware automatic chest X-ray radiology reports", or query a medical chatbot (12, 19, 26).

Unstructured (free-text) data contains domain-specific information which can be missed by structured fields of the EHR. The notes are usually recorded by nurses, and are a “highly untapped resource in clinical support” (40, 48). In a study that used a time-series, multi-modal approach for three ICU management related prediction tasks: in-hospital mortality, decompensation, and length of stay forecasting, it was found that adding clinical notes as another modality improves the performance of the model (23). Only patients with clinical notes who were admitted to the ICU for

48 hours were considered, and a convolutional neural network (CNN) was used to extract the textual features (23). The best-performance across all the tasks was achieved for the multimodal model using the convolutional approach, compared to the baseline (no notes), text only, and multimodal average word embedding without CNN (23).

Marafino et al. (2018) used data from January 1, 2001, through June 1, 2017 contributed by 20 ICUs at two academic medical centers and one community hospital in the United States. The patient study included the first ICU admission with a length of stay (LOS) of at least four hours, and used measures of clinical trajectory with clinical note features for in-hospital mortality prediction task (30). Their findings showed that incorporating variables measuring clinical trajectory, NLP-derived terms or both improved the model discrimination, and also demonstrated the "external validity and portability of models incorporating these variables" (30).

Weng et al. (2017) developed a ML-based NLP pipeline for the medical subdomain classification of clinical notes on the Integrating Data for Analysis, Anonymization, and Sharing (iDASH) and Massachusetts General Hospital (MGH) datasets, and found the best performance was achieved using a convolutional RNN and neural word embeddings with good transferability on the datasets (52). A bag-of-words representation was used with Apache clinical Text Analysis and Knowledge Extraction System (cTAKES), the Unified Medical Language System (UMLS) Metathesaurus, and learning algorithms to extract the features (52). cTAKES is a NLP-based, open source tool for medical notes annotation and information extraction from electronic health record clinical free-text (44).

2.4 Word embedding models

Word embeddings, or word vectors, are distributed representations of words that fall into two categories: frequency-based and prediction-based embeddings. Word2vec (32) is an example of a frequency-based embedding which uses negative sampling and allows transfer learning to take place. Transfer learning is mainly used in deep learning applications, and is when the knowledge gained when solving a problem can be applied to a different problem. Negative sampling is a technique that was developed to reduce computational cost by updating a small percentage of the language model's weights, resulting in an improved quality of the word vectors. FastText (21) is an efficient tool to learn word representations that uses n-gram characters, and generates better word embeddings for rare words compared to word2vec. The fastText architecture allows it to be used as an initializer for transfer learning as it supports continuous bag of words (CBOW) and skip-gram models (21).

BioWordVec (54) is an open set of pre-trained distributed word embeddings

based on text sequences from PubMed abstracts and clinical notes from the MIMIC-III dataset. BioWordVec employs the fastText subword embedding model and has been shown to have improved representations of rare biomedical terms in different NLP tasks in the biomedical domain(54). Hence, we decided to use the word embeddings from the BioWordVec model to explore whether using pre-trained embeddings shows a significant improvement in the model performance.

Bidirectional Encoder Representations from Transformers (BERT) (11) is a recent language model developed by Google which is used to obtain pre-trained deep bidirectional representations from unlabeled text without requiring "task-specific architecture modifications" (11). Transformers have the advantage of learning long-range dependencies, and uses self-attention to parallelize the computation and the bidirectional mechanism allows for the context to be incorporated from the left and right sides. Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT) is a contextual word representation model derived from BERT that uses a fine-tuning approach to group similar medical concepts (18). Two empirical studies were conducted to study the model performance: the first study explored language modeling and clinical word similarity which mapped medical terms on a representation space, and the second study was a 30-day hospital readmission prediction task. The findings from the two empirical studies showed an improved performance using clinical word similarities compared to using word2vec and a 15% relative increase on recall at a fixed rate of false alarms in a 30-day hospital readmission prediction task (18).

Oleynik et al. (2019) conducted a study aimed to evaluate the impact of using pretrained embeddings in a clinical text classification task across shallow and deep classifiers. They found that shallow ML strategies outperformed deep learning methods on small imbalanced data (36). The XGBoost model uses the Gradient Boosting Machine (GBM) framework for supervised learning problems, and is one of the “better performing shallow learning models in machine learning competitions" (6). We will choose to evaluate this model in addition to other shallow models including logistic regression and multinomial Naïve Bayes in our analysis.

Recently, Liu et al. (2019) combined features obtained from NLP of clinical notes with physiological data for early prediction of septic shock using a XGBoost algorithm (27). They compared two methods for generating NLP features: co-occurence matrix using document-term matrix (DTM), and RNN and Global Vectors for Word Representation (GloVe) (39) word embeddings for feature selection (27). Ten data points were selected per patient for the 28 physiological variables to reduce the model complexity; through uniform sampling for non-shock and from 1-2 hour time interval prior to septic shock for shock patients (27). The NLP features were concatenated with the physiological variables, and the best performance was achieved

using NLP features generated using GloVe word embeddings although the features generated using DTM were "more readily interpretable than those produced by the deep learning-based method" (27).

2.5 Conclusion

The challenge with relying on machine learning based models (such as neural networks) in the medical domain is due to the inherent black-box nature of the models. However, there does not necessarily need to be a trade-off between accuracy and interpretability of a model. As we are interested in designing and developing a system that will be used in the medical domain, we will focus on using interpretable models (43). Previous studies have looked at the effect of using multimodal systems (23, 52), and unlike different look back intervals with multivariate time-series data (45, 47) but they have not looked at the combination of these factors. We wish to take advantage of the time-series data by using the time-lagged embedding for the physiological factors in our model.

Moreover, unlike previous work we chose the 1992 sepsis definition (3) because it has a greater overlap with the factors in our selected BSPM model compared to the Sepsis-3 definition. We will focus on obtaining the mutual information between a set of note variables in combination with physiological time-series data to do a performance comparison with previous work.

Chapter 3

Methods

In this chapter, we present our sepsis prediction model and a description of the preprocessing steps and the algorithms used for training the model.

The tools used in this study were PostgreSQL[1] (using the PgAdmin 4 graphical interface), and Python 3.8 (50). The data cleaning and experimentation were run on a computer with an Intel Xeon W-2123 processor running at 3.60 GHz with 64.0 GB of RAM, running Windows 10.

3.1 Dataset

Our study uses the publicly available MIMIC-III[2] (v1.4) dataset, that includes EHR information for patients admitted to the ICU at the Beth Israel Deaconess Medical Centre in Boston, USA (20). The database consists of 26 linked tables and includes patient demographics, vital sign measurements, laboratory test results, procedures, medications, imaging reports, mortality (including post-hospital discharge), and 2 million clinical notes (20). The MIMIC-III dataset uses the International Classification of Diseases, Ninth Revision (ICD-9) codes that are used to assign diagnosis codes for classifying diseases, and for billing and clinical purposes. The ICD-9 codes consist of three to five numbers with the first three numbers representing the disease category, the fourth number narrowing it to a specific disease, and the fifth number differentiating between disease variants (33).

The database contains information from two distinct critical care information systems: Philips CareVue and iMDSoft Metavision, which differ according to how the data is stored and the time period that patients in each system were recorded: between 2001-2008 in the CareVue system and 2008-2012 in the MetaVision system (20). For this study, we have chosen the admissions recorded in the MetaVision

[1] https://www.postgresql.org/
[2] https://mimic.physionet.org/

system because there is less missing data (better data quality), and it includes the variables needed for our model.

The relationships between the tables are defined by the primary key (admission ID), and the relationship tables[3] are used to access the data from the relevant tables in the MIMIC-III database using PostgreSQL. Further processing is done on Python, and is outlined in section 3.3.

3.2 Patient Cohort

The structured EHR data and clinical notes for adult acute care patients were selected from the MIMIC-III dataset. Several inclusion and exclusion criteria were applied to the dataset, as shown in Figure 3.1 below. We restricted our analysis to the MetaVision subset due to the more updated data collection compared to CareVue and to avoid mapping of the two disjoint coding systems (20, 53). Furthermore, only adult patients ($\geq$18 years old) were considered to reduce model variability that would be introduced if newborn sepsis cases were also considered (13).

As a six hour interval was chosen as the minimum interval for the look back, patients with less than six hours of observations before reaching the septic condition onset time were excluded from the study. Furthermore, patients with an initial diagnosis of sepsis, and patient observations recorded after sepsis was detected were excluded from our study. The final cohort size is 33,928 admissions, comprised of 262,188 observations of which there are 65,395 unique clinical notes. The dataset is split into the training, validation, and testing sets using a 80-10-10 nonrandom split of the time-series data. This helps to avoid random variations between the training, validation and testing sets (34). The validation set is used to optimize the hyperparameters (described further in section 3.6.1), and the trained model is evaluated for the different performance metrics on the held-out test set.

The latest observations with unique clinical notes are extracted for each admission. The final number of observations for the training, validation and testing collections are 51160, 2116, and 12119, respectively. The reason that the testing data collection is higher compared to the validation is because the balancing of the testing and validation datasets was done after the 80-10-10 split through random undersampling.

[3] Refer to https://mit-lcp.github.io/ mimic-schema-spy/relationships.html

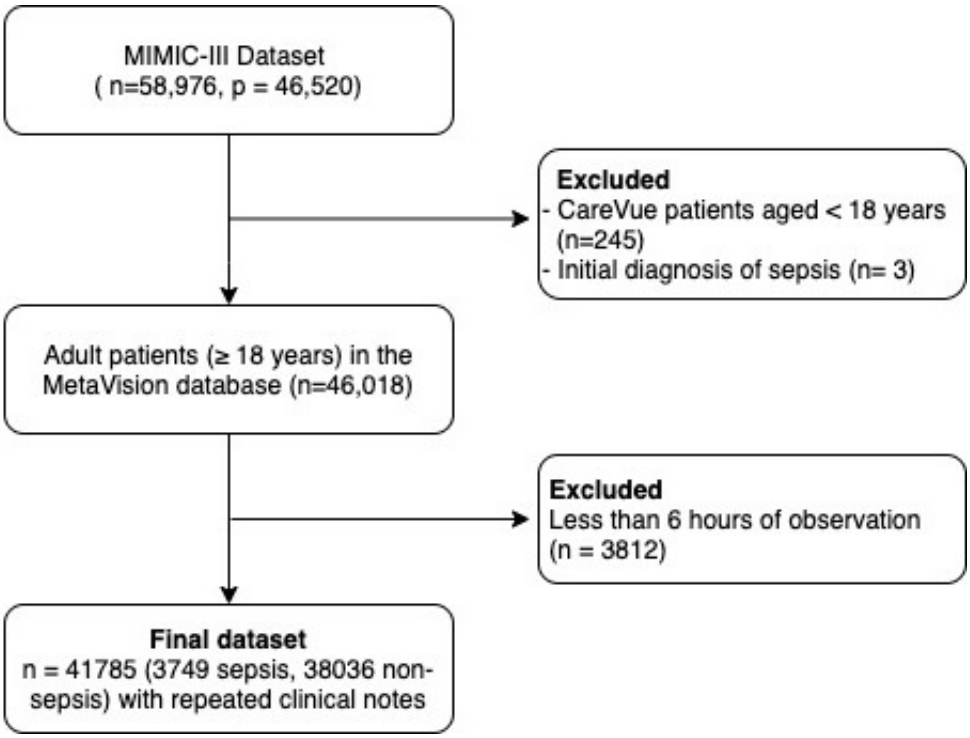

Figure 3.1: Admission inclusion chart with the selection criteria at each step, where p = unique patients, n = admissions. The chart shows the number of admissions that are included and excluded at each selection level. The final cohort size for the MIMIC-III dataset after preparation into 48 hour blocks is 41785 admissions, comprised of 3749 positive sepsis labels and 38036 negative labels

3.3 Data Preprocessing

The first step in successfully understanding the data was to identify the features used in related works (1, 45, 53), and that overlapped with the BSPM. The following libraries in Python were used for preprocessing the data: pandas (31) for working with large datasets, Scikit-learn (38) to code machine learning models, and NLTK (28) as a natural language toolkit for text processing. As a first step, initial exploratory data analysis was conducted using the `pandas profiling` tool to understand the data characteristics and distribution. The preprocessing steps for the physiological data imputation as outlined in (53) were followed.

Once the relevant information for each patient was extracted along with the time stamp for the lab tests and vital sign measurements, the medication count was recorded for the sepsis order set (described in section 2.1). The extracted time series data from MIMIC-III was re-sampled to hourly intervals with the last observation carried forward within each hospital admission stay for the missing vital signs. The remaining missing values were mean-imputed for each of the physiological factors (4).

We used two different data representations of the laboratory and vital measures for the baseline and the models that the experiments were run on. For the baseline model, the laboratory and vital measures were preprocessed for BSPM. The models that were developed for the experiments with time series and clinical notes used a second representation that took the deltas of the raw physiological measurements.

In order to compare the experiment results with the single time step experiments, a 48-hour time window was chosen for patient observations, given the maximum look back and prediction windows of 24 hours (53). The prediction time (PT) window is the time of the prediction ahead of sepsis onset. As a lower bound, only admissions with at least six observations were considered as a smaller window of observations would "have resulted in insufficient testing data" to make predictions, and it was also the shortest look back interval we considered (29). Figure 3.2 shows the feature windows for the sepsis and non-septic patients. For cases where there are less than 48 hours of observation for a patient, linear interpolation and “carry backward” extrapolation were done using the `interpolate()` function to standardize the input to 48-hour blocks.

The resulting admissions were balanced for the sepsis and non-septic patient admissions, and the number of admissions that satisfy this requirement has been shown in the admission inclusion chart in Figure 3.1. The differences between consecutive observations in the vital signs and laboratory measurements were computed, and transformed into time-series data taking a 24 hour look back for each observation. The remaining physiological factors are concatenated with the time-

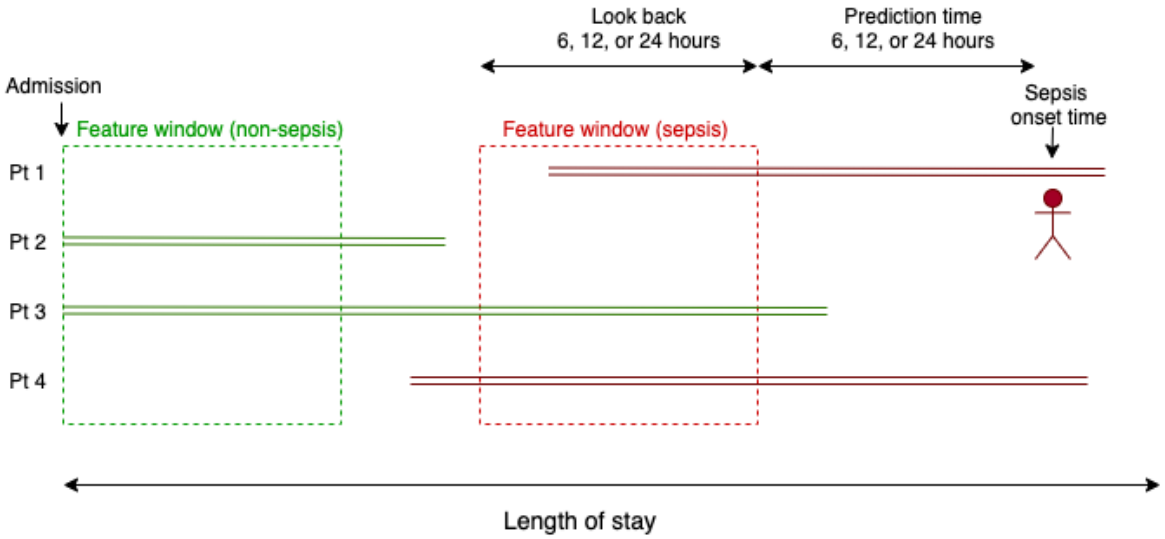

Figure 3.2: The model uses a combination of right and left aligned models for the sepsis (red) and non-septic (green) patients, respectively. For patients who do not develop sepsis for the duration of their stay, the feature window was taken as the initial 48 hours since admission. For the sepsis patients, diagnosis is made at a PT interval of 6, 12, or 24 hours ahead of sepsis onset time. The sepsis prediction will be positive if the sepsis onset time falls within the PT interval, and a minimum of 6 hours is considered for the look back interval which overlaps with the feature window

series variables, resulting in each admission having a 24 hour observation window. The output prediction labels for each observation depend on when sepsis onset occurs.

The detailed descriptions of the input and output variables are in sections 3.4 and 3.5 below.

3.4 Input variables

3.4.1 Physiological measures

For every admission in the study, we extracted a set of 38 variables from the MIMIC-III dataset, including demographics, vital signs, laboratory values, comorbidities, medications, procedural variables, and patient notes. The four SIRS conditions (including vital signs measurements and white blood cell (WBC)) were denoted as Pulse_score, Resp_score, Temp_score, and WBC_score, resulting in a total number of 54 physiological factors as input features to the model. A binary 'sus-

Table 3.1: List of ICD-9 codes related to the suspicion of infection in a patient

Description	ICD code
Septicemia	038
Septicemic, bacteremia, disseminated fungal/candida infection	0202, 7907, 1179, 1125
Disseminated fungal endocarditis	11281

pected_infection' variable was extracted for each patient admission using the ICD-9 diagnosis codes. Table 3.1 defines the ICD-9 codes that were used to define suspicion of infection in a patient.

Each observation corresponds to the time that a new vital chart signal measurement is recorded, and the remaining measurements were extracted according to their respective history windows up to the vital chart time, shown in Table 3.2. The missing lab values are left blank to indicate that they are not taken, while the other variables have the last observation carried forward. The medications were obtained from the `orderid` and `linkorderid` columns in the `inputevents_mv` table.

Table 3.2: This table illustrates the structured EHR data and clinical notes with the corresponding history windows from which they are extracted. The history windows were chosen based on related works and is expected to be sufficient for an accurate early prediction of sepsis

Category	History window
Vital sign measurements	All during patient stay
Laboratory results	Valid labs in previous 3 days
Comorbidities	All during the stay
Medications	1 day
Procedural variables	Currently active during stay
Patient notes	Latest note in previous 12 hours

3.4.2 Clinical Notes

The clinical notes were extracted from the `noteevents` table in the MIMIC-III database for the aforementioned cohort. As the observations were limited to patients who also had clinical notes, the performance between the physiological measures only model and the model incorporating the clinical notes are compared while maintaining the same train-validate-test split as before.

As the latest note is taken for each patient admission, a 12 hour interval was expected to be sufficient for ensuring that there is at least one note, and carry forward imputation was done for missing notes. Furthermore, discharge notes were

excluded from our sample because the early prediction of sepsis is made before the patient has been identified as septic, and before they are discharged from the hospital. The clinical notes are first preprocessed according to the steps used in Huang et al.'s paper, including text standardization, removing numbers, and fixing common misspellings (18). Additional preprocessing steps taken for the clinical notes involved removing section headers, compiling and excluding common stop words, tokenization, and encoding or vectorization (23). The final list of stopwords is comprised of NLTK's list of English stopwords combined with additional stopwords that were selected based on a bag-of-words model approach.

A more detailed description of the methodology pipeline is shown in Figure 3.3 below.

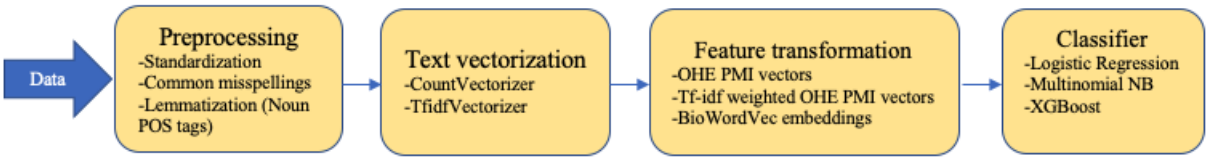

Figure 3.3: The graph illustrates the methodology pipeline for the clinical notes. The preprocessing involves text standardization, replacing common misspellings with correct spellings, and removing special characters. Next, text vectorization is done and followed by one of the three feature transformations: (1) one-hot encoding (OHE) of the PMI vectors, (2) multiplying the OHE PMI vectors by the Tf-idf weights, and (3) using pre-trained embedding vectors. Finally, the different models are run on the training data

Below is an example of the input text after it has been standardized and common misspellings have been replaced.

> *resp care pt remains intubated and currently vented on full support with changes made accordingly per abgs esophageal balloon . measurements obtained this shift and transpulmonary end exp pressure noted at at cmh peep . esophageal pressure cmh . bs dim course no sxing done this shfit . ett retaped due to massive edema around the face . peak plateau pressures respectively . pt remains metabolically acidotic and severely hypoxic despite vent changes . pt is dnr . will cont with vent support as needed .*

NLTK's `pos_tag` was used to obtain the part-of-speech tags, and the plural nouns (represented with 'NNS' and 'NNPS') were lemmatized into their singular

form. Furthermore, words with less than three letters are removed from the list of tokens as they are either stopwords or do not carry useful information.

As seen from the text below which shows the result of processing the input text, punctuation and prepositions including stopwords and words with less than three letters have been removed, leaving only the lemmatized form of the plural nouns.

> *resp care remains intubated currently vented full support chang made accordingly per abgs esophageal balloon measurement obtained shift transpulmonary end exp pressure noted cmh esophageal pressure cmh dim course sxing done shfit ett retaped due massive edema around face peak plateau pressur respectively remains metabolically acidotic severely hypoxic despite vent chang dnr cont vent support needed*

Three approaches were proposed for representing the notes as suitable input to the model, which involved doing text vectorization and feature transformation for the latest note at each time step, as described in the following sections.

The final note embeddings will be used as input to the model for the notes-only configuration, and is concatenated with the physiological factors for the multimodal setting before being fed into the different classifiers. The note embedding dimensions were chosen as 200 to make it comparable to the 200-dimensional pre-trained BioWordVec embedding vectors. Additionally, models with 200-300 dimensions are shown to have similar performance to models with larger embedding dimensions, as the dimensions should be chosen based on corpus statistics (37).

A. Pointwise Mutual Information (PMI) Matrix

The first approach uses pointwise mutual information (PMI) matrix, that is a measure of association between a feature and a class category to determine the association between the two words. This statistical measure developed by Church and Hanks (7) indicates how much the probability of a co-occurrence of events p(x,y) is different from the individual probabilities, and is represented by the following equation:

$$pmi(x;y) \equiv log\frac{p(x,y)}{p(x)p(y)}$$

This method is used to obtain the most important words relevant to sepsis prediction, and to select better features to model. The notes for a single admission are initially grouped together and labeled with ‘non-septic’ or ‘sepsis’ in order to extract the most relevant words that are associated with positive sepsis or negative sepsis. Once this is done, the following values are evaluated using the `CountVectorizer` and the PMI score for each (word, label) pair is calculated.

- C_{w+}: the number of sepsis admissions that contain word w
- C_{w-}: the number of non-septic admissions that contain word w
- C_{+}: the number of sepsis admissions
- C_{-}: the number of non-septic admissions
- C: the total number of admissions, i.e. sum(C_{+}, C_{-})
- C_{w}: the total number of admissions that contain w, i.e. sum(C_{w+}, C_{w-})

The PMI measure is obtained using the following formula for the sepsis admissions to find the top correlated words with sepsis.

$$pmi(word; class) \equiv log \frac{p(word, class)}{p(word)p(class)}$$

$$pmi(w; +) \equiv log \frac{\frac{C_{w+}}{C}}{\frac{C_{+}}{C} \cdot \frac{C_{w}}{C}}$$

$$pmi(w; +) \equiv log \frac{C_{w+} \cdot C}{C_{+} \cdot C_{w}}$$

Only the relevant terms are kept by taking the words with top 200 positive PMI scores, and one-hot encoding is performed for each of the unique notes. The resulting note embeddings are the note features that will be used as input to the model for the notes-only configuration, and concatenated with the physiological factors obtained earlier for the multimodal setting.

B. Term frequency-inverse document frequency PMI method

The second embedding method uses a TfidfVectorizer which tokenizes the text and assigns weights to the pre-selected PMI words according to the importance of the word in a document. `TfidfVectorizer` is used to get the Tf-idf weights for the one-hot encoded PMI words from the previous method, and terms that occur more frequently in one document compared to the rest of the corpus gets assigned higher importance. The minimum number of times that the word should appear in the corpus is represented by `min_df` and it was set to 5. The average embedding vector per document was calculated by taking the dot product of the Tf-idf vector and the one-hot encoded vectors obtained in the previous step.

C. BioWordVec pre-trained embeddings

The third document embedding method uses the distributed representation of words through the pre-trained word embedding vectors available from BioWordVec (54). The BioWordVec model is trained on PubMed abstracts and clinical notes in the MIMIC-III dataset. FastText (21) was applied to compute 200-dimensional word embeddings with the following settings: window size = 20, learning rate = 0.05, sampling threshold = 1e-4, and negative examples = 10. Sentence vectors are obtained for each of the clinical notes by calling the `get_sentence_vector(sent)` function, where `sent` represents the preprocessed note.

The following functions can be called from the fasttext model:
`model.get_word_vector(word)`: gets the word representation, and
`model.get_input_vector(word_ID)`: gets the word representation with the given word ID.

3.5 Output variables

A patient is determined to have sepsis if their ICD-9 code is consistent with suspected infection, and two out of four SIRS criteria shown in Table 2.1 are satisfied (3). Three output variables with a binary value o $\in$ {0, 1} are defined for each PT interval of 6, 12, and 24 hours, as whether the patient has contracted sepsis within the given PT window.

This means that the output labels for patients who develop sepsis during their ICU stay will be zero until the sepsis onset occurs within their PT interval.

3.6 Model development

We prepared data collections to explore the effect of using different history of the variable (using different look back intervals) on the performance of the sepsis onset prediction model, by concatenating each row with (n-1) previous observations, where n $\in$ {6, 12, 24}. As shown in Figure 3.2, a combination of left and right aligned models are used to represent the non-septic and sepsis patients. The methodology by Hsu et al. (17) was followed to model the temporal classification problem which was used to assess the sepsis onset risk score on the MIMIC-III dataset.

The baseline performance was evaluated by running the model on the BSPM, and was compared with the models with physiological (time series) features, note features alone, and the combination of both to see which gives the best performance

compared to the BSPM performance. The evaluation metrics are listed in Chapter 4 and include the commonly used c-statistic, or AUROC metric, and the Area under the Precision-Recall Curve (AUPRC) metric which is useful when the dataset is highly-skewed (10).

A description of the models used is included below:

- **Logistic regression:** Multivariable logistic regression methods are used to explore the risk factors associated. Logistic regression models are used to estimate the relationships between a dependent variable and one or more independent variables, or predictors. Implementing a logistic regression model helps to achieve a better understanding of how the model works, as it is easy to interpret and trains quickly on the dataset.

- **XGBoost (XGB):** XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework to find an optimal model that fits the data. Decision Trees do not require normalization of their inputs, thus the unscaled data can be passed to the model. Furthermore, XGBoost is useful in feature selection since we are able to obtain the feature importance of each factor. At each step during the hyperparameter tuning, the AUROC and precision scores were calculated for the training and validation sets. The number of decision trees (N_estimators) was set to 100, and the following hyperparameters were tuned sequentially/pairwise at each step:

  - max_depth: the size of the decision trees. There is a trade-off between shallow trees (or weak learners) and deeper trees (tends to overfit).
  - min_child_weight: the minimum sum of weights of all observations required in a child, and tuning this parameter helps to control over-fitting.
  - gamma (default=0): specifies the minimum loss reduction required to make a further partition.
  - subsample: the fraction of observations to be randomly samples for each tree.
  - colsample_bytree: the fraction of columns randomly sampled for each tree.
  - regularization alpha: L1 regularization term on weight

- **Multinomial Naïve Bayes (MNB):** This is a commonly used parametric model for text categorization problems due to its computational advantage as it only considers words with a non-zero count. MNB maximizes the likelihood over the accuracy.

The clinical note embeddings are incorporated with the physiological factors at the **feature-level** by concatenating the 200-dimensional note embeddings with the time series variables and training a single model. Standardization and scaling the magnitude of each variable is important for the above models, to reduce the effect of large magnitudes which may have an undesired effect on the final prediction. `MinMaxScaler()` is an estimator that scales each feature to a given range, which is usually between zero and one. This scaler is used to scale the data before running the the logistic regression and multinomial Naïve Bayes classifiers. The XGB model does not require scaling as it can handle unscaled input.

3.6.1 Hyperparameters

The model is trained with using a stratified split on the full dataset using 80% as the training set, 10% as a validation set to select the hyperparameters, and a final 10% for testing (17). The training and validation sets are balanced for the sepsis and non sepsis cases while maintaining the same splits over time.

For the LR model, we searched for the regularization parameter using ParameterGrid, $C \in 10^{[-3,2]}$ on a logarithmic scale. For the MNB model, we searched for $alpha \in [0.5, 1.5]$ on a linear scale, with $fit\_prior \in [True, False]$. For the XGB model, a 5-fold cross validation is used to select the best values for the hyperparameters using GridSearchCV (38). The number of estimators was fixed at 100, and the settings for the XGB model were obtained in multiple steps. Firstly, we searched for the $max\_depth \in [3, 10]$ and $min\_child\_weight \in [1, 6]$. Once the best parameters were found, we used those values and searched for gamma $\in [0, 0.5]$. Similarly as before, we used the best parameter and searched for $subsample \in [0.6, 1]$, and $colsample\_bytree \in [0.6, 1]$ on a linear scale. Finally, we searched for the $reg\_alpha \in [1e-5, 1e-2, 0.1, 1, 100]$.

Once all the hyperparameters are chosen, we ran the model on the test set and recorded the performance metrics which have been compiled in Chapter 4.

Chapter 4

Results

In this chapter, we present the performance metrics of the different models. Initial descriptive analytics were first conducted using Python's `pandas profiling` tool to better understand the dataset, and the figures are included in section 4.1. Section 4.2 outlines the evaluation metrics that were used in this study.

Section 4.3 presents the results of running the models on the structured physiological measures: single time data preprocessed similar to the benchmark model and time-series data. Section 4.4 shows the results for the notes-only model and Section 4.5 shows the results from running the models on the combined physiological time series and notes data according to the three different note embedding methods described in the previous chapter.

4.1 Initial Data Analysis

Since the MIMIC-III dataset preparation involved shifting the dates of birth for patients older than 89 to preserve anonymity, those ages were shifted back to the realistic age range, while maintaining the median age as 91.4 (20). This was done by subtracting 200 from all the ages that were originally shifted.

Afterwards the data is split into the training, validation and testing sets which are normalised using the MinMaxScaler from the `sklearn` library, and mean-imputed for the missing values before the models are run. There are 12 distinct note categories in the original MIMIC-III database, and the top five categories make up more than 90% of the total notes. The 'discharge summary' notes were not considered in our experiment as they are usually recorded after the patient has been discharged from the hospital and thus, can not be used to predict sepsis onset in a patient. Table 4.1 below shows the different note categories and their counts. The final training-validation-testing sample had 65,395 unique doctor notes.

The description of the patient cohort included in the balanced training and valida-

Table 4.1: Overview of the top 5 note categories and their frequencies in the MIMIC-III database

Note category	Count (%)
Nursing/other	240537 (54.6)
Radiology	140366 (31.9)
Nursing	27890 (6.3)
Physician	20401 (4.6)
Respiratory	7063 (1.6)

.

tions sets are presented in Table 4.2. The training set was composed of 56.8% men with a mean age of 63 years. Of the 22748 patient admissions, 7432 were found to have sepsis (21.8%). For the validation set, 243 were found to have sepsis (36.0%).

Table 4.2: Description of the patient cohort included in the training and validation sets used to train and develop our system

Demographics	MIMIC-III training set	MIMIC-III validation set
Number of admissions	34092	675
Number of observations	145713	16191
Males (%)	19364 (56.8)	381 (56.4)
Mean age	63.19	63.52
Sepsis condition (%)	7432 (21.8)	243 (36.0)

4.2 Evaluation Metrics

We evaluated the model performance using the following metrics:

Area Under the Receiver Operating Characteristic curve (AUROC)

The Area Under the ROC curve (AUROC) summarizes the performance of a model in terms of its ROC curve and ranges from 0.5 (no skill) to 1.0 (perfect classification). The ROC curve shows the relationship between clinical sensitivity and specificity and plots the sensitivity on the y-axis and (1-specificity) on the x-axis. Sensitivity is the ability of a test to correctly identify patients with the disease. Improving the sensitivity of a test means reducing the false negative results and minimizing missing cases of disease.

$$Sensitivity = \frac{TP}{TP + FN}$$

where TN: true negative, TP: true positive, FN: false negative, and FP: false positive. Specificity is the ability of the test to correctly identify patients without the disease and is represented by the following:

$$Specificity = \frac{TN}{TN + FP}$$

Area Under the Precision-Recall Curve (AUPRC)

The Area Under the Precision-Recall Curve (AUPRC) is a useful performance metric when dealing with imbalanced data where proper classification of the positive cases is important. The 'average precision' method is used to calculate the AUPRC and the baseline is equal to the fraction of positive examples.

4.3 Physiological-based models

The performance of different machine learning models were evaluated and compared with the baseline model. Only the observations with unique clinical notes are considered, and the missing input physiological features are imputed as described in subsection 3.4.1 in chapter 3. Hyperparameter optimization is done on the validation set using grid search parameter tuning, and the chosen hyperparameters are run on the test set for all of the models. The hyperparameters for each setting are given in the last column, and the boldfaced numbers in each column of the table indicate the highest performance metrics for the given dataset.

4.3.1 Results of Baseline model: single time-instant

The baselines are defined for the logistic regression, multinomial Naïve Bayes, and XGBoost models as the physiological measures preprocessed similar to the BSPM, with each row corresponding to EHR data from a single time step. The physiological measures have been transformed using MinMaxScaler from a range of 0 to 1 for the logistic regression and MNB models for improved performance. The XGB model does not require normalization or missing value imputation since it is done natively by the model.

Table 4.3 summarizes the performance metrics of the baseline model which has the input physiological features preprocessed for BSPM. As seen from the table, the highest performing model is the XGB_{BASE} model which achieves an AUROC of 0.78 and AUPRC of 0.30. The LR_{BASE} has the next highest AUROC score, but achieves a lower AUPRC compared to the MNB_{BASE} model that has the lowest AUROC.

The precision-recall and receiver operating characteristic (ROC) curves for the logistic regression, MNB, and XGB models, respectively are shown in Figure A.1 in Appendix A. The ROC curves

Table 4.3: Results from running the baseline models for the physiological features preprocessed for BSPM. The XGB model achieves the best performance in terms of the AUROC and AUPRC, compared to the other models. There is a directly proportional relationship between the model performance and the PT intervals

Model	Prediction time (hours)	AUROC	AUPRC	Hyperparameters
LR_{BASE}	6	0.709	0.098	C=0.001
	12	0.722	0.144	C=0.001
	24	0.733	0.182	C=0.001
MNB_{BASE}	6	0.711	0.113	alpha=1.3, fit_prior=T
	12	0.718	0.160	alpha=1.3, fit_prior=T
	24	0.724	0.201	alpha=0.7, fit_prior=T
XGB_{BASE}	6	0.745	0.123	max_depth=3, min_child_weight=5, gamma= 0.2, subsample= 0.6, colsample_bytree=0.6, reg_alpha= 100
	12	0.760	0.212	max_depth=3, min_child_weight=3, gamma=0.4, subsample=0.9, colsample_bytree=0.6, reg_alpha=0.1
	24	**0.780**	**0.303**	max_depth=3, min_child_weight=5, gamma=0, subsample=0.6, colsample_bytree=0.9, reg_alpha=1e-5

4.3.2 Time-series data models

In this section, we present the results of running the models on the physiological measures represented as time series input data to study whether it achieves an improvement in performance compared to a single time point.

Table A.1 shows the results of running the model on the time series input measures. The XGB model performed best out of the three models, and the LR model achieved the worst performance in both metrics. While the MNB and XGB models' performance is better compared to the single time instant, the LR model performs worse except for an improved performance for the 12 hour PT. This can be observed from the precision-recall and ROC curves for the logistic regression and MNB models shown in Figure A.2.

4.4 Text-based models

Different clinical notes representations were compared to identify which embedding approach gives the best performance. We compared three approaches to get the document embedding; the first two made use of the top 200 PMI words (unweighted and weighted by Tf-idf), and the third utilized the pre-trained embedding vectors from the BioWordVec model. The mapping of the notes into the corresponding document embeddings is described in section 3.4.2. The different models are then run on the input note features and the results of running the model on these features are presented below. The boldfaced numbers in each column of the table indicate the highest performance metrics for the given dataset.

Pointwise Mutual Information (PMI) one-hot encoded embeddings

The first document embedding we explored was the PMI one-hot encoded embedding. The 200 most informative words related to sepsis with the top 200 PMI scores were obtained, and the notes were represented by one-hot encoded vectors.

Table 4.4 shows the performance results for the models trained on the note features. The XGB model achieves the highest performance overall, with 0.769 AUROC, 0.298 AUPRC for the 24 hour PT interval. It is followed by the LR model, and the MNB model has the lowest performance with 0.628 AUROC, 0.197 AUPRC for the 24 hour PT interval.

Table 4.4: Results from running the notes-only model with the notes represented as one-hot encoded PMI vectors with dimensions n=200. The XGB model achieves the highest AUROC and AUPRC scores, followed by the LR model, and the MNB achieves the lowest scores

Model	Prediction time (hours)	AUROC	AUPRC	Hyperparameters
LR	6	0.731	0.175	C=0.01
	12	0.736	0.219	C=0.1
	24	0.746	0.252	C=0.01
MNB	6	0.622	0.141	alpha=1.3, fit_prior=T
	12	0.628	0.167	alpha=0.5, fit_prior=T
	24	0.628	0.197	alpha=1.1, fit_prior=T
XGB	6	0.751	0.191	max_depth=9, min_child_weight=5, gamma=0.3, subsample=0.8, colsample_bytree=0.8, reg_alpha=1
	12	0.760	0.25	max_depth=9, min_child_weight=5, gamma=0.2, subsample=0.8, colsample_bytree=0.7, reg_alpha=1e-5
	24	**0.769**	**0.298**	max_depth=9, min_child_weight=3, gamma=0, subsample=0.9, colsample_bytree=0.6, reg_alpha=0.01,

Tf-idf weighted PMI one-hot encoded embeddings

The 200 one-hot encoded vectors obtained in the previous embedding are element-wise multiplied by the Tf-idf weights. Table 4.5 shows the results of running the model on the note features, and Figure A.3 show the precision-recall and ROC curves for the LR, MNB, and XGB models.

Table 4.5: Results of running the model with the notes represented by 200-dimensional tf-idf weighted PMI embeddings. The XGB model achieves the highest AUROC and AUPRC scores, which are comparable with the scores of the LR model. The MNB model performs the worst, for all PT intervals

Model	Prediction time (PT)	AUROC	AUPRC	Hyperparameters
LR	6 hr	0.666	0.093	C=0.001
	12 hr	0.666	0.107	C=0.001
	24 hr	0.673	0.126	C=0.001
MNB	6 hr	0.650	0.095	alpha=0.5, fit_prior=T
	12 hr	0.648	0.108	alpha=0.5, fit_prior=T
	24 hr	0.652	0.128	alpha=0.5, fit_prior=T
XGB	6 hr	0.741	0.147	max_depth=1, min_child_weight=1, gamma= 0, subsample=0.9, colsample_bytree=0.6, reg_alpha=1
	12 hr	0.740	0.187	max_depth=7, min_child_weight=3, gamma=0, subsample=0.6, colsample_bytree=0.9, reg_alpha=1e-05
	24 hr	**0.757**	**0.226**	max_depth=7, min_child_weight=3, gamma=0, subsample=0.6, colsample_bytree=0.6, reg_alpha=1

A comparison with Table 4.4 shows that for the 24 hour PT interval, the unweighted PMI vectors have a higher performance compared to the Tf-idf weighted PMI vectors. The ROC and PR plots are similar to each other, with only the XGB model showing an evident difference in the area under the curve for longer PT intervals.

BioWordVec embeddings (n=200)

Finally, the results of running the model on the clinical notes using BioWordVec 200-dimensional pre-trained embeddings are given in Table A.3. As seen from the table, the highest performance is observed with the longest PT of 24 hours, for the XGB model.

4.5 Multimodal models

The final stage of the experiment is to combine the physiological time series variables with the note features encoded with the three document encoding approaches. As in the previous section, we compare the performance of the PMI one-hot encoding with PMI weighted by Tf-idf with the pre-trained embeddings obtained using the BioWordVec model.

PMI one-hot encoded embeddings

Table 4.6 shows the summary of the results of running the model on the combined EHR features with the notes represented as PMI one-hot encoded embeddings. We can observe that the performance metrics for the multimodal model are higher compared to the physiological-based model (refer Table 4.3) as well as the note embeddings model (Table 4.4). This is in accordance with our hypothesis, that adding note embeddings to the physiological measures will result in an improved prediction. The XGB model achieves the best performance in terms of the AUROC and AUPRC, followed by the LR model, and lastly the MNB model.

Figure A.4 shows the ROC and precision-recall curves for the combined time series physiological input with PMI one-hot embeddings. As seen from the plots, the area under the curve is highest for the XGB model, followed by the LR model. The MNB model has the least area under the precision-recall and ROC curves, with lower variation for the different prediction times. Since the latter has not achieved better performance in the experiments thus far compared to XGB and LR, we decided to exclude the MNB model from future iterations.

Table 4.6: Results of running the model on the combined time series physiological measures and notes represented as one-hot encoded PMI vectors with dimensions n=200. The XGB model achieves the best performance in terms of the AUROC and AUPRC, compared to the LR and MNB models. We can see that the multimodal model performs better than the physiological measures model and the PMI encoded notes only model

Model	Prediction time (hours)	AUROC	AUPRC	Hyperparameters
LR	6	0.805	0.235	C = 0.1
	12	0.805	0.287	C = 0.1
	24	0.812	0.336	C = 0.1
MNB	6	0.750	0.154	alpha=0.5, fit_prior=F
	12	0.750	0.180	alpha=0.5, fit_prior=T
	24	0.755	0.215	alpha=0.5, fit_prior=F
XGB	6	0.826	0.253	max_depth=9, min_child_weight=5, gamma=0.3, subsample=0.6, colsample_bytree=0.6, reg_alpha=.1
	12	0.831	0.338	max_depth=9, min_child_weight=5, gamma=0, subsample=0.6, colsample_bytree=0.7, reg_alpha=1
	24	**0.834**	**0.384**	max_depth=9, min_child_weight=3, gamma=0, subsample=0.9, colsample_bytree=0.6, reg_alpha=1e-5,

Tf-idf weighted PMI one-hot encoded embeddings

Next, the notes are encoded with Tf-idf weighted one-hot encoded PMI vectors. Figure A.5 shows the precision-recall and ROC curves for the LR and MNB graphs. The results of the model performance are presented in Table A.2 in Appendix A. We observe that the performance of the model trained with the combined features are consistently better compared to the notes-only and and physiological time series models. The Tf-idf weighted notes-only model has a higher AUROC across the models compared to the physiological data only.

BioWordVec embeddings (n=200)

Finally, the results of running the model on the combined physiological and notes features obtained using the BioWordVec embeddings for a 24 hour look back interval are given in Table A.4 in Appendix A. The XGB model achieves the best performance, followed by the LR model, and then the MNB model. While the XGB and LR models' performances are higher for the combined BioWordVec embedding compared to the combined Tf-idf weighted PMI encodings, the MNB performance is

comparable (ranging from 0.75-0.77).

4.5.1 Comparing different look back intervals

All the models that we have seen so far (for the single time step and time series data) have used a 24 hour look back interval. Focusing on the time series representation, a further comparison between the different look back time intervals (6, 12 and 24 hour) was done using the best-performing document embedding to represent the clinical notes. The results of running the logistic regression and XGB models on the different look back intervals are shown in Tables 4.7 and 4.8, respectively.

From Table 4.7, we can observe that the LR model achieved the best performance for the 24 hour look back interval, and the longest PT interval. A closer look shows that the AUROC values for the 6 and 12 hour look back intervals are approximately the same, with the AUPRC score slightly higher for the 6 hour look back. This means that there are more false alerts, and less precise detection for the 12 hour look back compared to the 6 hour interval.

Looking at Table 4.8 for the XGB model, a similar trend is observed with the XGB model achieving the best performance for the 24 hour look back, and the 24 hour PT interval. The AUROC and AUPRC values are similar for the 6 and 12 hour look back intervals, with the latter being slightly better with higher AUPRC scores.

Table 4.7: Comparison of different look back times (6, 12 and 24 hours) of the Logistic Regression model with the combined factors as time series input, using BioWordVec embedding representation

Look back (hr)	Prediction time (hr)	AUROC	AUPRC	Hyperparameters
6	6	0.691	0.105	C= 0.001
	12	0.781	0.214	C = 100
	24	0.783	0.241	C = 100
12	6	0.691	0.105	C = 0.001
	12	0.782	0.208	C = 100
	24	0.784	0.236	C = 100
24	6	0.828	0.194	C = 100
	12	0.843	0.266	C = 1
	24	**0.855**	**0.331**	C = 1

Table 4.8: Comparison of different look back times (6, 12 and 24 hour) of the XGB model with the combined factors as time series input, using BioWordVec embedding representation

Look back (hr)	Prediction time (hr)	AUROC	AUPRC	Hyperparameters
6	6	0.810	0.23	max_depth=3, min_child_weight=5, gamma=0, subsample=0.8, colsample_bytree=0.9, reg_alpha=1
	12	0.812	0.248	max_depth=3, min_child_weight=3, gamma=0, subsample=0.9, colsample_bytree=0.7, reg_alpha=0.1
	24	0.818	0.293	max_depth=5, min_child_weight=1, gamma=0.4, subsample=0.8, colsample_bytree=0.8, reg_alpha=1
12	6	0.808	0.216	max_depth=3, min_child_weight=5, gamma=0.3, subsample=0.8, colsample_bytree=0.8, reg_alpha=1e-5
	12	0.82	0.256	max_depth=5, min_child_weight=3, gamma=0, subsample=0.9, colsample_bytree=0.6, reg_alpha=0.01
	24	0.827	0.306	max_depth=5, min_child_weight=1, gamma=0.4, subsample=0.8, colsample_bytree=0.8, reg_alpha=1
24	6	0.841	0.201	max_depth=9, min_child_weight=5, gamma=0.1, subsample=0.9, colsample_bytree=0.6, reg_alpha=1e-5
	12	0.854	0.28	max_depth=9, min_child_weight=3, gamma=0, subsample=0.8, colsample_bytree=0.8, reg_alpha=1e-5
	24	**0.878**	**0.400**	max_depth=7, min_child_weight=5, gamma=0.4, subsample=0.8, colsample_bytree=0.6, reg_alpha=1

The chart in Figure 4.1 shows the comparison of the AUROC scores of the logistic regression and XGB models for the combined time series physiological variables and notes represented by the BioWordVec embeddings, for the 24 hour PT interval. We observe that the LR notes-only model outperforms the 6 and 12 hour look back intervals in terms of the AUROC, while the XGB notes-only is comparable to the same intervals. The 24 hour look back interval achieves the best performance in terms of the AUROC for both models.

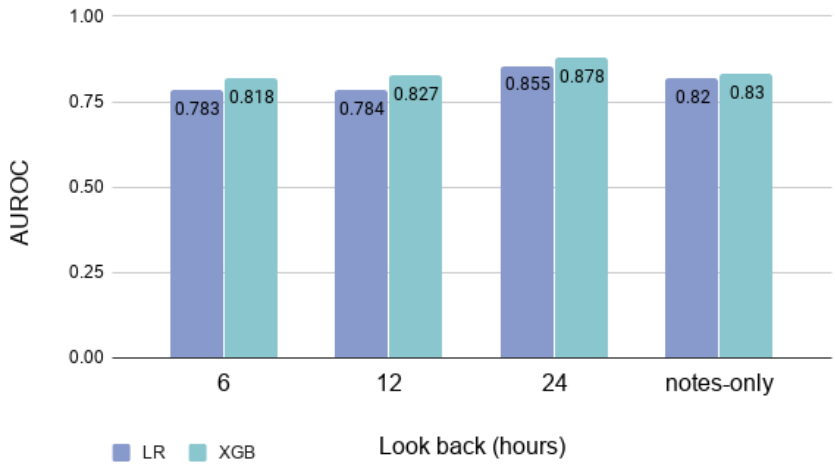

Figure 4.1: The chart compares the AUROC of the LR and XGB models for the combined physiological features and notes encoded using BioWordVec embeddings, for a PT interval of 24 hours. The notes-only model is compared with the combined features for the 6, 12, and 24 hour look back intervals. The highest AUROC is achieved for the longest look back interval by the XGB model followed by the LR model

4.6 Summary

In this chapter we have looked at the different combinations of single versus time series data, notes-only and multimodal models for three document embedding methods, and the effect of different look back intervals.

Table 4.9 provides a summary of the performance of the physiological-based, text-based and multimodal models with the different note embeddings for the best performing models in each category. The multimodal models achieve the highest performance out of them all, followed by the text-based which implies that the note features captured useful information for sepsis onset prediction. The XGB model achieves the best performance for the Tf-idf weighted PMI one-hot encoding and the BioWordVec embeddings.

The tuned hyperparameters of the XGB notes-only model were used for the combined model as it did not change much for a given PT as seen from tables 4.4 and 4.6.

Figure 4.1 illustrates the differences for the logistic regression and XGB models across the different look back intervals for the 24 hour PT interval. As seen in the figure, the highest AUROC is observed for the longest (24 hour) look back interval. Interestingly, the notes-only model achieves the next highest AUROC score for the LR model, and very similar values for the XGB model (notes-only: 0.83, 12 hour look

Table 4.9: Performance comparison of the physiological-based model, text-based model, and multimodal model with 24 hour PT for the XGB model which achieves the best-performance with the different note embedding representations. The highest AUROC and AUPRC are obtained for the XGB model using BioWordVec embedding, for the 24 hour PT

Models	**Document embedding**	**AUROC**	**AUPRC**
Physiological-only	-	0.816	0.352
Notes-only	PMI one-hot encoding	0.769	0.298
	Tf-idf weighted PMI one-hot encoding	0.740	0.187
	BioWordVec embedding	0.830	0.355
Combined	PMI one-hot encoding	0.834	0.384
	Tf-idf weighted PMI one-hot encoding	0.855	0.366
	BioWordVec embedding	**0.878**	**0.400**

back: 0.827). This finding suggests that the NLP features for the LR model includes terms that are predictive features for sepsis onset.

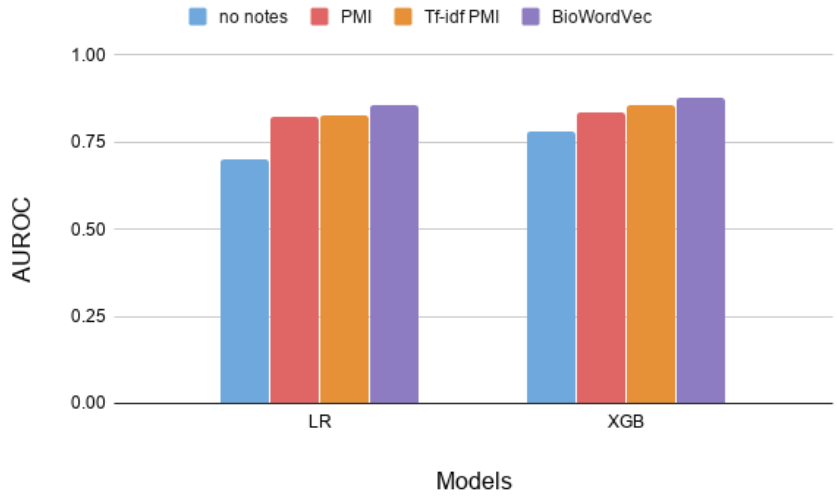

Figure 4.2: Comparison of the AUROC scores for the LR and XGB models using the three note-embedding approaches, with a 24 hour PT. The BioWordVec embedding for the notes gives the best performance, then the Tf-idf weighted one-hot encoded PMI vectors, then one-hot encoded PMI vectors. We observe that the AUROC scores for the latter two encodings are very similar for the LR model

Figure 4.2 shows a comparison of the AUROC scores for the LR and XGB models using the three note-embedding approaches with a 24 hour PT interval. The document embedding that made use of the pre-trained embedding vectors from the BioWordVec model performed the best overall, with an AUROC of 87.8% for a 24 hour look back window and 24 hour PT interval.

Furthermore, the AUROC scores of the best performing model are in the range of 0.78-0.9 which indicate an acceptable to excellent model discrimination.

Chapter 5

Discussion

In this chapter, we present the discussion of the results from our experiments from the previous chapter. Our findings support the research hypothesis that models incorporating the note embeddings perform better than models taking either the physiological or note features alone as input. We evaluated logistic regression, multinomial Naïve Bayes, and XGBoost (XGB) models on three configurations: a) structured data components of the EHR, b) clinical note embeddings alone, and c) the combination of structured EHR and note features.

5.1 Prediction Method

The sepsis onset prediction task was framed as a supervised learning problem with pre-defined labels based on the Sepsis-2 definition. The Sepsis-2 definition was chosen based on the factors that were present in the BSPM, and future studies could consider using feature engineering methods such as principal component analysis (PCA) and linear discriminant analysis (LDA) to reduce the model complexity.

We can see how different sets of features perform on the prediction task by comparing the performance of the models for the given prediction time (PT) and look back intervals (6, 12, and 24 hours). Firstly, it was found that longer look back and PT windows result in a higher performance compared to the baseline model in terms of the AUROC and AUPRC scores across the different settings. A fixed interval of the most recent 48 hours was extracted for each patient: before sepsis onset for sepsis patients and the first 48 hours for non-septic patients (53). This was chosen to reflect the the real life scenario a longer patient history may not be accessible for the look back, nor for the PT intervals.

Secondly, the performance of the baseline model is higher than the note-only features, but lower than the combined features which suggests that the note features alone might not be a good feature set compared to the physiological mea-

sures alone. The combined features always results in an improved prediction for all the embeddings. While the BioWordVec embedding gives the highest performance overall for the document embeddings, the Tf-idf weighted PMI encoding is a close second and may be preferred in cases where the clinicians or doctors wants to know the specific words (or concepts) that influenced the prediction. As an alternative to our topic modeling approach of using PMI to extract the thematic words related to sepsis, we may also use LDA to further improve the model accuracy.

Furthermore, another observation relating to the text-based models is that the Tf-idf weighted encodings achieved lower performance compared to the unweighted PMI encodings for the LR and XGB models, and similar performance for the multimodal models as seen from Figure 4.2.

5.2 Acceptability in hospitals

Systems that provide clinicians with evidence-based recommendations are implemented in hospitals, such as the AI Pathway Companion at the University Hospital Basel (USB) that uses NLP on radiology and pathology results to provide patient-specific risk assessments (51). However, the main challenges facing the full-fledged implementation of AI systems in medical decision-making is the black-box modeling nature and the lack of interpretability of those systems.

Efforts need to be taken to involve the clinicians in the design and development of these systems, and incorporate their feedback following the human-in-the-loop approach. Further integration into hospital settings could be done with the use of voice technology and automated text recognition tools which would allow for more textual information to be recorded and be made available.

The governance of these systems should be done by a committee of different stakeholders including local policymakers, clinicians, and AI manufacturers to ensure that accountability and ethical principles are followed. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines should be followed in publishing and reporting any clinical machine learning models (8).

The early sepsis prediction tool should be used as an aide for clinicians rather than a standalone device, and it's important that the patients are aware of the fact that these technologies are being used, and clinicians should receive their explicit consent before using these systems (41).

Chapter 6

Conclusion and Future Works

In this thesis, we have illustrated the different parts of developing an early sepsis prediction model that incorporates clinical notes with structured EHR data to improve performance. This includes data extraction and processing, defining and formulating the research questions in Chapter 1, and designing experiments to test our hypothesis. Three different approaches for encoding the clinical notes are presented for three machine learning models: linear regression, multinomial Naïve Bayes, and XGB. This chapter summarises the contribution of our research in section 6.1, and the future work in section 6.2.

6.1 Conclusion

Our main research question is *whether a more accurate sepsis prediction model can be obtained by incorporating clinical notes with time-series physiological measures*. A comparison of the performance metrics indicates that combining the physiological and note features (in multimodal models) results in the highest performance compared to the disjoint features. In some cases, the performance of the notes only model is higher than the physiological features model indicating that important information for the sepsis onset prediction task is captured by the note features.

The second research question addresses *which clinical note representations and shallow ML models are best suited to predict sepsis onset in adult ICU patients*, and found that the XGBoost model performed better across the different configurations, followed by the LR model. The multinomial naïve Bayes had the lowest performance overall. Furthermore, the BioWordVec embeddings achieved better performance compared to the other document embedding approaches.

We further explored *the effect of changing the look back and prediction time (PT) intervals on the model performance*. It was found that the longer look back and PT intervals results in a higher performance AUROC and AUPRC scores across

the different settings. In real life scenarios, however, a longer patient history may not be accessible for the look back intervals. The XGB model attained the highest performance of 0.88 AUROC, 0.40 AUPRC with the multimodal features and the notes encoded with BioWordVec embeddings for a 24 hour look back and prediction time intervals.

6.2 Future Work

The Sepsis-2 definition was chosen to define sepsis in this study instead of the Sepsis-3 definition from 2016 because it had a greater overlap with the factors present in the BSPM (49). Future studies could compare the performance of the model while using the Sepsis-3 definition which uses the SOFA score, which may be a better discriminant than the traditional SIRS which has a lack of specificity and increased sensitivity (2, 46).

Future iterations of the sepsis prediction model may use neural network models such as LSTM networks or RNN which take advantage of memory when dealing with time-varying input and could be used to uncover hidden factors that lead to sepsis (22). This should be done while explaining why the algorithm gives the prediction either through incorporating attention mechanisms into the models, or performing post-hoc feature importance analysis using the tree SHapley Additive exPlanations (SHAP) (35).

The current approach of obtaining the PMI features for the clinical note encoding employs a BOW-based approach. An alternative to using a statistics-based approach is to use a pattern-based modeling approach that can help to recognize the features and procedures from the clinical notes which is useful to find out what kind of features are more prominent in predicting sepsis onset (29).

The research could be improved upon by exploring the use of ensemble models, where the notes and the physiological features are trained separately at the classifier-level and then aggregated to predict sepsis onset. The loss of the data when no clinical notes are available can also be addressed by implementing a **notes_on** mode which may be toggled during the running of the prediction model: ‘on’ when clinical notes are available and ‘off’ when not available.

Finally, numerous models were developed and tested on the freely-available MIMIC-III database which comprises of patients of a particular demographic (4, 23, 27, 29, 45, 53). It is important to carry out a multicentre validation study to ensure generalizability as the choice of the training data may result in implicit bias that favors people from certain countries, demographics, or age groups.

Bibliography

[1] Armando D Bedoya, Joseph Futoma, Meredith E Clement, Kristin Corey, Nathan Brajer, Anthony Lin, Morgan G Simons, Michael Gao, Marshall Nichols, Suresh Balu, Katherine Heller, Mark Sendak, and Cara O'Brien. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open*, 3(2):252–260, 04 2020.

[2] Tony Berger, Jeffrey Green, Timothy Horeczko, Yolanda Hagar, Nidhi Garg, Alison Suarez, Edward A Panacek, and Nathan Shapiro. Shock index and early recognition of sepsis in the emergency department: Pilot study. volume 14, pages 168–174. University of California, 2013.

[3] Roger C. Bone, Robert A. Balk, Frank B. Cerra, R. Phillip Dellinger, Alan M. Fein, William A. Knaus, Roland M.H. Schein, and William J. Sibbald. Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis. *CHEST*, 101(6):1644–1655, Jun 1992.

[4] Jacob S. Calvert, Daniel A. Price, Uli K. Chettipally, Christopher W. Barton, Mitchell D. Feldman, Jana L. Hoffman, Melissa Jay, and Ritankar Das. A computational approach to early sepsis detection. *Computers in Biology and Medicine*, 74:69 – 73, 2016.

[5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[6] F. Chollet and J.J. Allaire. *Deep Learning with R*. Manning Publications, 2018.

[7] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March 1990.

[8] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G.M. Moons. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). *Circulation*, 131(2):211–219, 2015.

[9] Irene Cortes Puch and Christiane Hartog. Change is not necessarily progress: Revision of the sepsis definition should be based on new scientific insights. *American journal of respiratory and critical care medicine*, 194, 05 2016.

[10] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML '06*, 2006.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 2019.

[12] Rashmi Dharwadkar and Neeta A. Deshpande. A Medical ChatBot. *International Journal of Computer Trends and Technology (IJCTT)*, 60(1):41–45, 2018.

[13] Mahmoud ElHalik, Javed Habibullah, Khaled El-Atawi, and Amany Abdelsamad. Epidemiology of Sepsis in NICU; A 12 Years Study from Dubai, U.A.E. *Journal of Pediatrics & Neonatal Care*, 8(2):259–72, 2018.

[14] C Fleischmann, A Scherag, and NK Adhikari. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. *Am J Respir Critical Care Med*, 193(3):259–72, 2016.

[15] Lucas Fleuren, Thomas Klausch, Charlotte Zwager, Linda Schoonmade, Tingjie Guo, Luca Roggeveen, Eleonora Swart, Armand Girbes, Patrick Thoral, Ari Ercole, Mark Hoogendoorn, and Paul Elbers. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, 46, 01 2020.

[16] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

[17] Po-Ya Hsu and Chester Holtz. A comparison of machine learning tools for early prediction of sepsis from icu data. 2019.

[18] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 2019.

[19] Richard G Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, and Robert Stewart. Natural Language Processing to Extract Symptoms of Severe Mental Illness from Clinical Text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open*, 7(1), 2017.

[20] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. FastText.zip: Compressing text classification models. 2016.

[22] Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S Dangayach, Anthony B Costa, and Eric K Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PloS one*, 14(2):e0211057, 2019.

[23] Swaraj Khadanga, Karan Aggarwal, Shafiq R. Joty, and Jaideep Srivastava. Using Clinical Notes with Time Series Data for ICU Management. In *EMNLP/IJCNLP*, 2019.

[24] Anand Kumar, Daniel Roberts, Kenneth Wood, Bruce Light, Joseph Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, David Gurka, Aseem Kumar, and Mary Cheang. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine*, 34:1589–96, 07 2006.

[25] S. Lambden, P. Laterre, M. Levy, and B. Francois. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23, 2019.

[26] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR.

[27] R. Liu, J. L. Greenstein, S. Sarma, and R. Winslow. Natural language processing of clinical notes for improved early prediction of septic shock in the icu. pages 6103–6108, 2019.

[28] Edward Loper and Steven Bird. Nltk: The natural language toolkit. page 63–70, 2002.

[29] Qingqing Mao, Melissa Jay, Jana L Hoffman, Jacob Calvert, Christopher Barton, David Shimabukuro, Lisa Shieh, Uli Chettipally, Grant Fletcher, Yaniv Kerem, Yifan Zhou, and Ritankar Das. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. *BMJ Open*, 8(1), 2018.

[30] Ben J. Marafino, Miran Park, Jason M. Davies, Robert Thombley, Harold S. Luft, David C. Sing, Dhruv S. Kazi, Colette DeJong, W. John Boscardin, Mitzi L. Dean, and R. Adams Dudley. Validation of Prediction Models for Critical Care Outcomes Using Natural Language Processing of Electronic Health Record Data. *JAMA Network Open*, 1(8):e185097–e185097, 12 2018.

[31] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[33] Elias Moons, Aditya Khanna, Abbas Akkasi, and Marie-Francine Moens. A comparison of deep learning methods for icd coding of clinical records. *Applied Sciences*, 10(15):5262, 2020.

[34] Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyenberg, Andrew J. Vickers, David Ransohoff, and Gary S. Collins. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanantion and Elaboration. *Annals of Internal Medicine*, 162(1):W1–W74, 2015.

[35] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew Stanley, Gari Clifford, and Timothy Buchman. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical Care Medicine*, 46:1, 12 2017.

[36] Michel Oleynik, Amila Kugic, Zdenko Kasáč, and Markus Kreuzthaler. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association : JAMIA*, 26, 09 2019.

[37] Kevin Patel and Pushpak Bhattacharyya. Towards lower bounds on number of dimensions for word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*,

pages 31–36, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[40] Philip Resnik, Michael Niv, Michael Nossal, Andrew Kapit, and Richard Toren. Communication of clinically relevant information in electronic health records : A comparison between structured data and unrestricted physician language. 2008.

[41] Rebecca Robbins and Erin Brodwin. Patients aren't being told about the AI systems advising their care, Jul 2020.

[42] Kristina E. Rudd, Sarah Charlotte Johnson, Kareha M. Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V. Colombara, Kevin S. Ikuta, Niranjan Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R. Machado, Konrad K. Reinhart, Kathryn Rowan, Christopher W. Seymour, R. Scott Watson, T. Eoin West, Fatima Marinho, Simon I. Hay, Rafael Lozano, Alan D. Lopez, Derek C. Angus, Christopher J. L. Murray, and Mohsen Naghavi. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, Jan 2020.

[43] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2018.

[44] G. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, S. Sohn, K. Schuler, and C. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17 5:507–13, 2010.

[45] Matthieu Scherpf, Felix Gräßer, Hagen Malberg, and Sebastian Zaunseder. Predicting Sepsis With a Recurrent Neural Network Using the MIMIC-III Database. *Comput Biol Med.*, 113(103395), 2019.

[46] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762–774, 02 2016.

[47] Supreeth P Shashikumar, Qiao Li, Gari D Clifford, and Shamim Nemati. Multiscale network representation of physiological time series for early prediction of sepsis. *Physiological Measurement*, 38(12):2235–2248, nov 2017.

[48] Benjamin Shickel, P. Tighe, A. Bihorac, and Parisa Rashidi. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22:1589–1604, 2018.

[49] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810, 02 2016.

[50] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[51] Philipp Grätzel von Grätz. Clinical decision support for prostate cancer care, 2020. `https://www.siemens-healthineers.com/news/ai-pathway-cds-basel.html/`, Last accessed on 2020-09-09.

[52] Wei-Hung Weng, Kavishwar B. Wagholikar, Alexa T. McCray, Peter Szolovits, and Henry C. Chueh. Medical Subdomain Classification of Clinical Notes Using a Machine Learning-Based Natural Language Processing Approach. *BMC Medical Informatics and Decision Making.*, 17, 2017. doi:`10.1186/s12911-017-0556-8`.

[53] Ke Yu, Mingda Zhang, Tianyi Cui, and Milos Hauskrecht. Monitoring ICU Mortality Risk with A Long Short-Term Memory Recurrent Neural Network. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25:103 – 114, 2020.

[54] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Bioword-vec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019.

Appendix A

Additional Results

The tables, ROC curves, and precision-recall curves for the different configurations from the Results section are presented in this section.

Table A.1: Results from running the model on time series physiological measures. The XGB model achieved the best performance in terms of the AUROC and AUPRC, and the LR model performed the worst. An interesting observation is that AUROC and AUPRC of the LR model for the 12 hour PT is highest for the same model and compared to the MNB models

Model	Prediction time (hours)	AUROC	AUPRC	Hyperparameters
LR	6	0.691	0.095	C=0.001
	12	0.781	0.192	C=10
	24	0.699	0.143	C=0.001
MNB	6	0.714	0.118	alpha=1.5, fit_prior=T
	12	0.719	0.155	alpha=1.5, fit_prior=T
	24	0.726	0.188	alpha=1.5, fit_prior=T
XGB	6	0.767	0.136	max_depth=3, min_child_weight=5, gamma= 0, subsample=0.6, colsample_bytree=0.7, reg_alpha=1e-5
	12	0.786	0.226	max_depth=3, min_child_weight=5, gamma=0, subsample=0.6, colsample_bytree=0.9, reg_alpha=1
	24	**0.816**	**0.352**	max_depth=5, min_child_weight=5, gamma=0.3, subsample=0.7, colsample_bytree=0.6, reg_alpha=0.1

Table A.2: Results from running the model on combined notes (tf-idf weighted PMI) and time series physiological measures, for a 24 hour look back interval. There is consistent improvement in performance compared to the unweighted PMI document encodings for all of the models, and the XGB model achieves the best performance followed by the LR model and the MNB model.

Model	Prediction time (hours)	AUROC	AUPRC	Hyperparameters
LR	6	0.809	0.254	C=1
	12	0.819	0.248	C=1
	24	0.827	0.299	C=1
MNB	6	0.750	0.127	alpha=0.9, fit_prior=T
	12	0.760	0.168	alpha=0.7, fit_prior=T
	24	0.769	0.213	alpha=1.1, fit_prior=T
XGB	6	0.823	0.175	max_depth=9, min_child_weight=5, gamma= 0, subsample=0.6, colsample_bytree=0.7, reg_alpha=1e-5
	12	0.849	0.267	max_depth=7, min_child_weight=3, gamma=0, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1
	24	**0.855**	**0.366**	max_depth=9, min_child_weight=5, gamma=0.1, subsample=0.6, colsample_bytree=0.8, reg_alpha=1

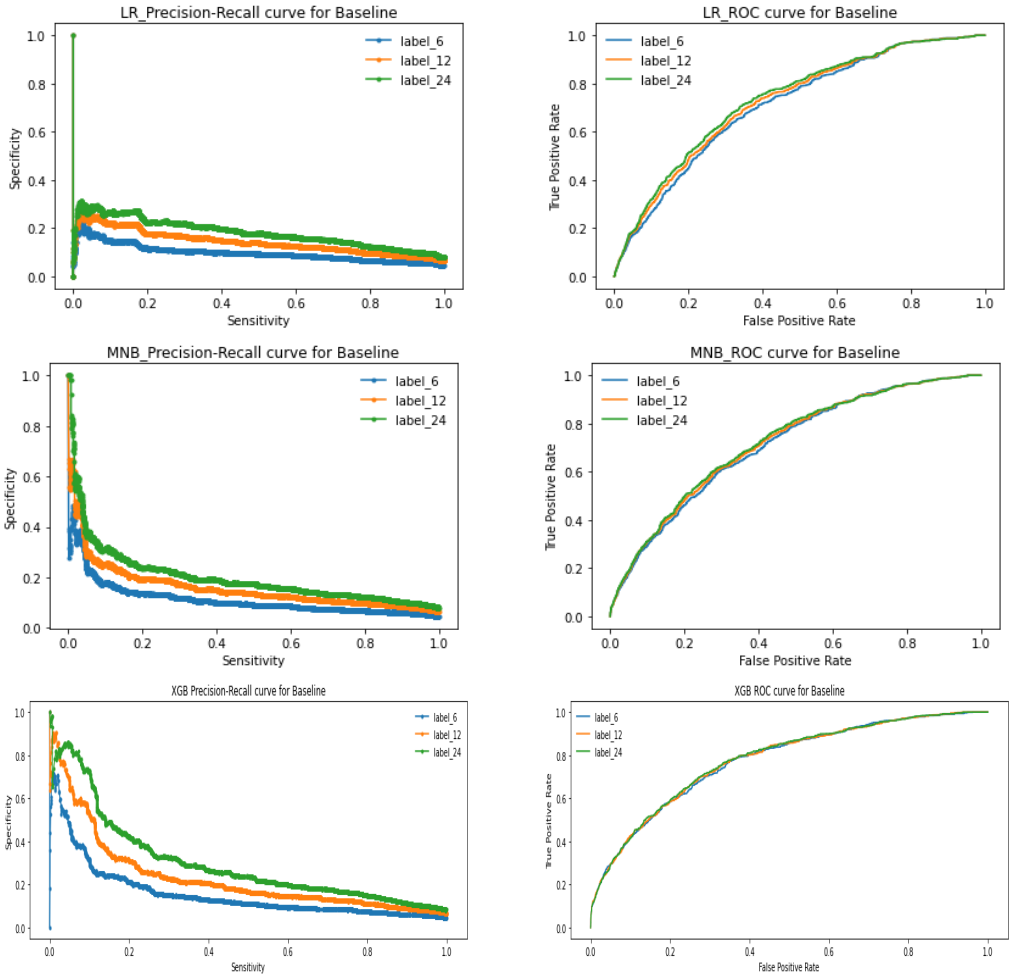

Figure A.1: Precision-recall and ROC curves for the baseline LR (top), MNB (middle), and XGB (bottom) models for 6, 12 and 24 hour PT interval. While the ROC curves are almost similar for the different PT intervals, we observe that the longer PT interval has a higher area under the curve across all the models

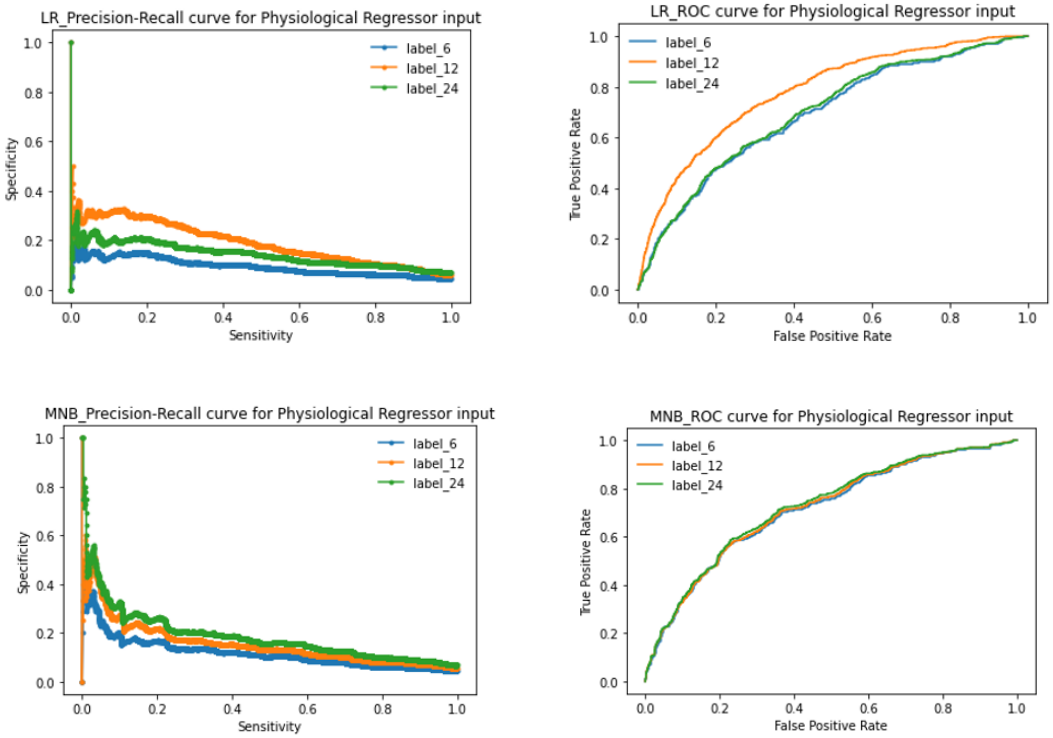

Figure A.2: Precision-recall and ROC curves for the physiological measures as time series input features for 6, 12 and 24 hour PT for the LR (top) and MNB (bottom) models. We can observe for the LR model that the 12 hour PT has a higher ROC curve (and AUROC score) compared to the other curves

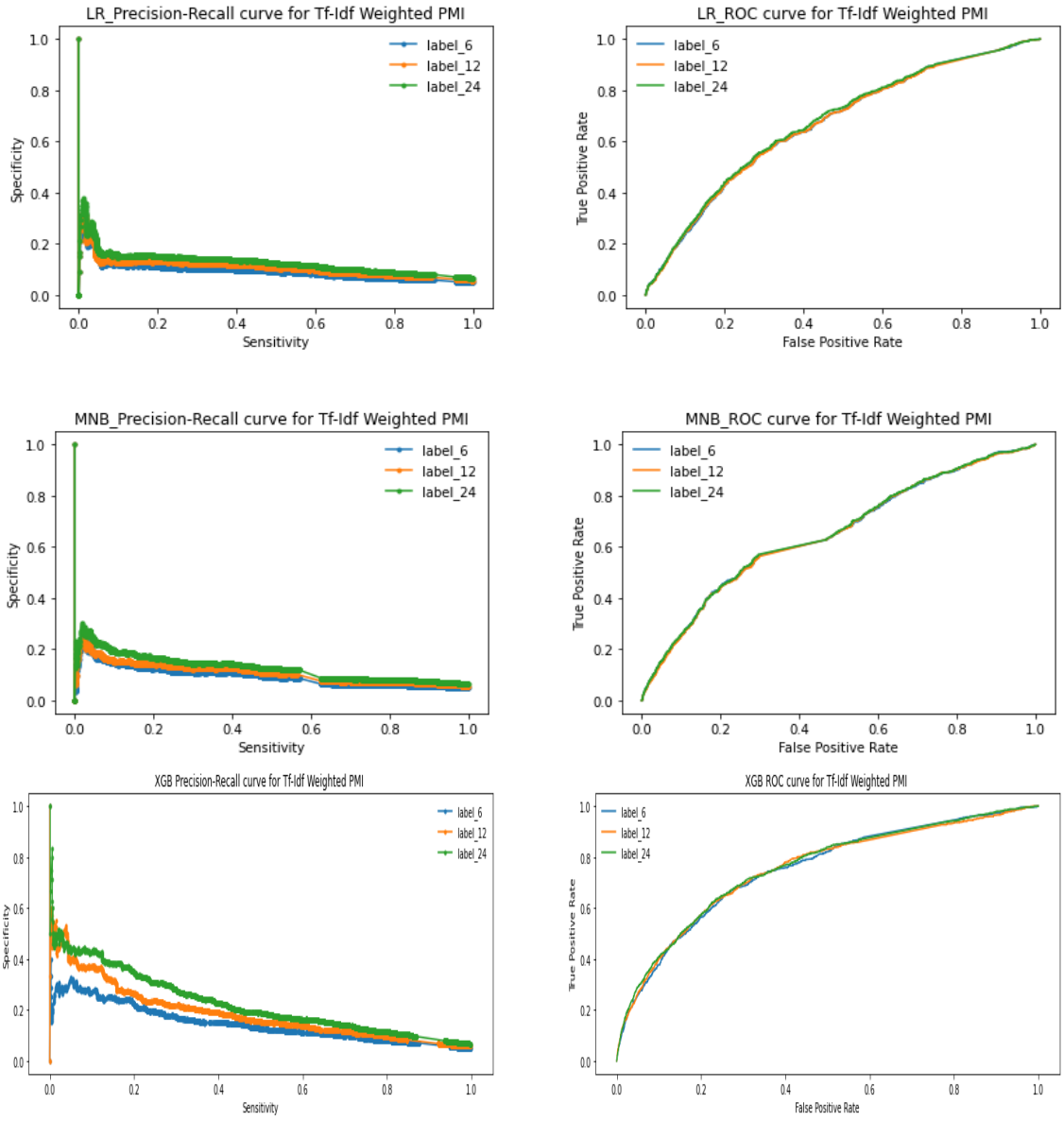

Figure A.3: Precision-recall and ROC curves for the Tf-idf weighted PMI note features for 6, 12 and 24 hour PT for the LR (top), MNB (middle), and XGB (bottom) models. The ROC curves are similar across all the models for all PT intervals, while the XGB model shows a noticeable increase in the area under the curve for longer PT intervals

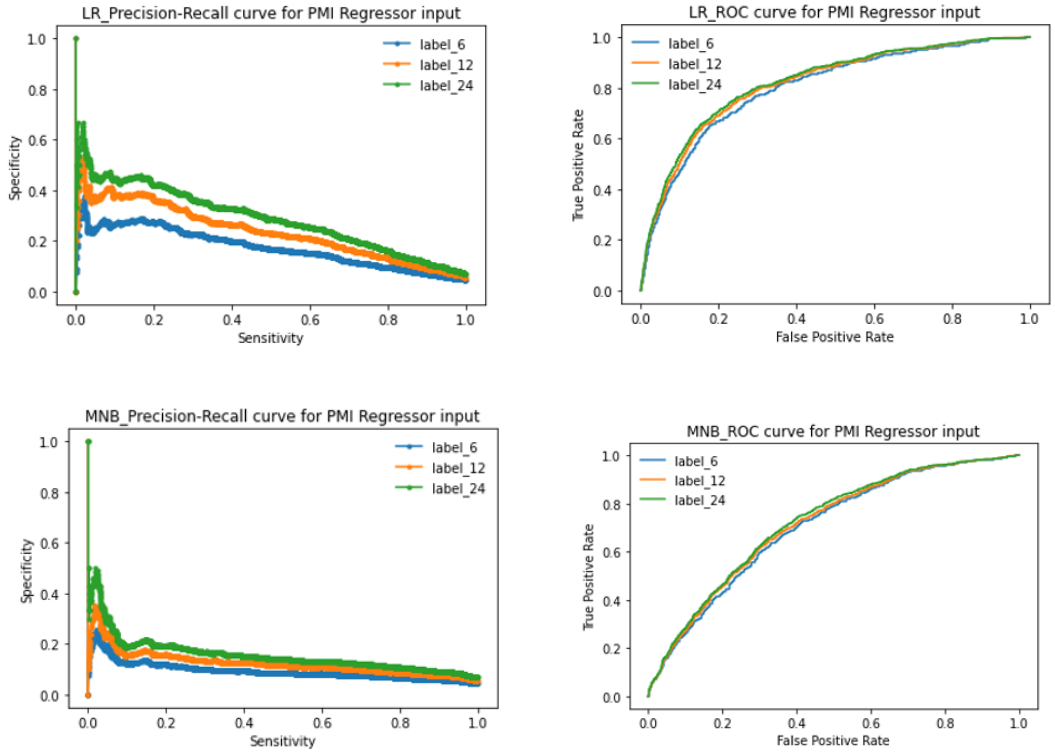

Figure A.4: Precision-recall and ROC curves for the time series physiological input features and PMI one-hot encoded notes for the LR (top) and MNB (bottom) models. The LR model achieves the best performance with the highest area under the curves, compared to the MNB model curves.

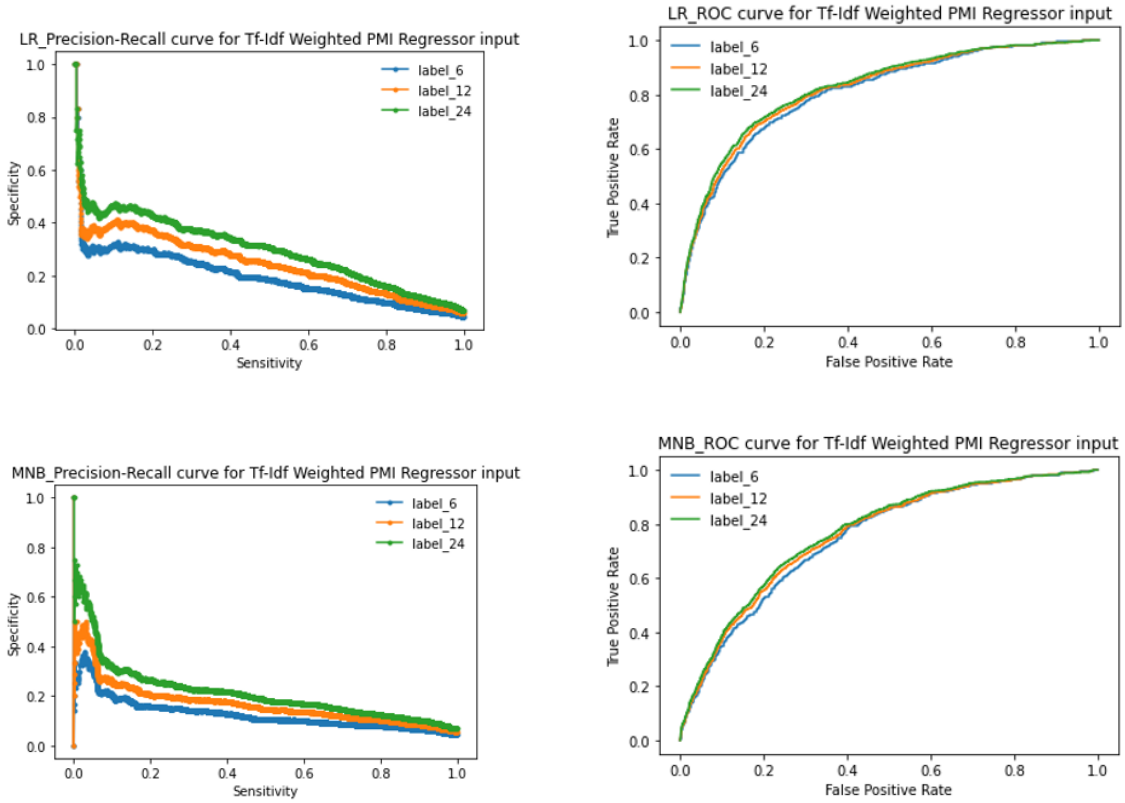

Figure A.5: Precision-recall and ROC curves for the combined time series input features and Tf-idf weighted PMI notes features for 6, 12 and 24 hour PT for the LR (top) and MNB (bottom) models. It is clear from the graphs that the longer PT interval has higher specificity for a given sensitivity. The ROC curves are similar for all PT intervals, with minor improvements for longer PT intervals

Table A.3: Results of running the model with the notes represented by 200-dimensional BioWordVec embeddings. The XGB model achieves the highest AUROC and AUPRC scores, which are comparable with the scores of the LR model. The MNB model achieved the worst performance, for all PT intervals

Model (tfidf)	Prediction time (hours)	AUROC	AUPRC	Hyperparameters
LR	6	0.807	0.200	C=10
	12	0.813	0.250	C=1
	24	0.820	0.293	C=1
MNB	6	0.686	0.110	alpha=0.5, fit_prior=T
	12	0.690	0.135	alpha=0.5, fit_prior=T
	24	0.694	0.157	alpha=0.5, fit_prior=T
XGB	6	0.807	0.230	max_depth=7, min_child_weight=3, gamma= 0, subsample=0.7 colsample_bytree=0.9 reg_alpha=0.01
	12	0.819	0.304	max_depth=9, min_child_weight=5, gamma=0.2, subsample=0.9, colsample_bytree=0.6, reg_alpha=0.1
	24	**0.830**	**0.355**	max_depth=7, min_child_weight=5, gamma=0, subsample=0.8, colsample_bytree=0.8, reg_alpha=1e-5

Table A.4: Results from running the model on the combined time series physiological measures and notes represented using the BioWordVec embeddings (n=200) for a 24 hour look back interval. The XGB model achieves the best performance as indicated by the bolded values, followed closely by the LR model. Furthermore, the improvements in the AUROC are incremental for the PT intervals

Model	Prediction time (hours)	AUROC	AUPRC	Hyperparameters
LR	6	0.837	0.254	C=1
	12	0.840	0.315	C=1
	24	0.846	0.367	C=1
MNB	6	0.761	0.165	alpha=0.5, fit_prior=T
	12	0.757	0.192	alpha=0.5, fit_prior=T
	24	0.765	0.235	alpha=0.5, fit_prior=T
XGB	6	0.841	0.201	max_depth=7, min_child_weight=3, gamma= 0, subsample=0.7 colsample_bytree=0.9, reg_alpha=0.01
	12	0.854	0.280	max_depth=9, min_child_weight=5, gamma=0.2, subsample=0.9, colsample_bytree=0.6, reg_alpha=0.1
	24	**0.878**	**0.400**	max_depth=7, min_child_weight=5, gamma=0, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1