Predicting Patient Outcomes from Biosignal and EHR Data

Team Members: Tegan Ayers, Mehdi Safari, Parv Thakkar, Hongyang (Jarvis) Wang

GitHub:

https://github.com/parv212/Predicting-Patient-Outcomes-from-Biosignal-and-EHR-Data

Summary:

Healthcare is a highly lucrative industry that could benefit significantly, both in regard to patient outcomes and cost savings, from machine learning and predictive modeling. While some progress has been made in this space, including image classification of tumors, natural language processing of electronic health records (EHR), and predictive models for cancer treatments, ample amounts of opportunities remain.

One challenge to developing highly accurate and generalizable models for healthcare is the lack of large, diverse, and time-aligned datasets. The creators of the VitalDB dataset [1] sought to address this challenge by aggregating information from EHRs and vital sign monitoring systems in order to create a rich dataset that contains highly granular data with pre-, intra- and postoperative patient outcomes. Specifically, datastreams include discrete features such as demographics, diagnoses and normative blood biomarker levels, as well as continuous biosignals such as heart rate (HR), body temperature and blood pressure, each collected at various sample rates.

The goal of this project is to develop four models that predict several patient outcomes, including patient diagnoses, surgical anesthesia duration, ICU length of stay and mortality rate using the VitalDB dataset. These target variables were selected because they are representative of a surgical patients entire journey throughout the hospital system, from being admitted to the hospital (i.e. diagnosis), completing surgery (i.e. anesthesia duration), and ultimately recovery time (i.e. ICU length of stay and mortality rate). The intention is to build models that only utilize predictors that were collected prior to that phase of the patient's journey, such that the next phase of the journey can be predicted prior to it occurring, thereby allowing the hospital system to more effectively plan resources.

During Phase 1, initial modeling was completed using the EHR data only, and base machine learning models such as Naive Bayes, Random Forest, XGBoost and fully-connected neural networks were implemented. Results are shown in Table 1.

Phase 2 consisted of improving upon the models in several ways. For the patient diagnosis model, classes were grouped in order to reduce the overall number of diagnoses and a multi-output model was generated which output the top 3 predictions along with confidences. These changes resulted in a 10% increase in accuracy for this response variable. In contrast, for the ICU duration and mortality models, the time series biosignal data was processed and used as inputs into the models. In addition, an LSTM model architecture was implemented for comparison. Unfortunately, these changes did not result in significant changes to model

performance and it is recommended to prioritize the simpler, Phase 1 models instead. Finally, the anesthesia duration model was dropped due to lack of model utility. (i.e. an anesthesiologist would likely do a better job predicting this outcome).

In summary, the outcome of this project was to develop three models that, when evaluated together, help the patient and hospital system be more informed of a patient's medical journey. While this was achieved, the models would likely need to undergo further improvements prior to real-world implementation. Specifically, more diverse data, especially in regards to patient race, should be included during training in order to make these models more generalizable to the world population.

Table 1: Phase 1 Summary of Results

Target Variable	Model	Performance
Patient Diagnosis	Random Forest	Accuracy: 36.2%
Anesthesia Duration	Lasso Regression	RMSE: 0.33
ICU Stay (Boolean)	XGBoost	Accuracy: 91%
ICU Duration (days)	Random Forest	Accuracy: 85%
Mortality Rate	XGBoost	Accuracy: 95%

Dataset:
The Vital Signs Database (VitalDB) [1] is an open source database. Data was obtained from non-cardiac surgery patients who underwent routine or emergency surgery at Seoul National University Hospital in South Korea from August 2016 to June 2017. The dataset contains clinical and biosignal data from 6,388 patients, including 73 electronic health record (EHR) parameters, 34 biomarker laboratory results and over 196 unique biosignal waveforms across all patients.

Methods:
Individual methods for each model are presented below.

Patient Diagnosis
For this model, only relevant EHR data was used as inputs to the model. Data cleaning included removing unnecessary columns and using only features that were relevant to patient diagnosis. This included removing any features that may have been collected *after* initial patient admission (e.g., ICU duration). Next, basic data preprocessing, such as specifying data types, label and one-hot encoding and handling missing values was performed.

Prior to cleaning, the total number of patient diagnoses was over 1,300, of which some only contained 1 or 2 samples. As such, any diagnoses containing less than 7 samples were removed. Then, the variable "operation type" was used to group the diagnoses, which ultimately

reduced the total number of classes to 10 and improved accuracy. Table 2 presents a sample of some of the groupings, which shows that the diagnoses were heavily grouped by body part. It's important to note that operation type was then removed as a feature from the dataset so as not to bias the model.

Table 2: Patient Diagnosis Grouped by Operation Type

Operation Type (Used to Group Classes)	**No. of Unique Diagnoses Present**
Colorectal	34
Breast	5
Biliary/Pancreas	20
Stomach	4

After cleaning, the dataset size was [4858, 56]. An 80%-20% train-test split was utilized during training and a multi-output, hyperparameter-optimized Random Forest model was trained and evaluated.

ICU Duration

Similar data cleaning steps as described above were used to clean the EHR data for this model, including removing unnecessary columns, handling missing values and label encoding. This resulted in an EHR dataset of size [6386, 1463]. In addition, the time series biosignal data was also included as inputs to this model. Given the team's limited domain knowledge of biosignals, only 3 of the 196 available biosignals were used. These included heart rate (HR), body temperature (BT), and bispectral index (BIS), which is a representation of anesthesia depth during surgery. These signals were also the most prevalent across all patients, with HR being contained within 100% of patient files, BT within 93% and BIS within 87%. Each signal was trimmed to the length of the surgery, then padded with zeros to be equal to the length of the longest surgery (15.675 hours or 28,215 samples). Signals were sampled every 2 seconds (0.03 Hz) and any missing values were replaced with a dummy variable (-999). This resulted in the biosignal data taking the shape [6386, 3, 28215]. Figure 1 represents the three signals without any padding added for ease of visualization.

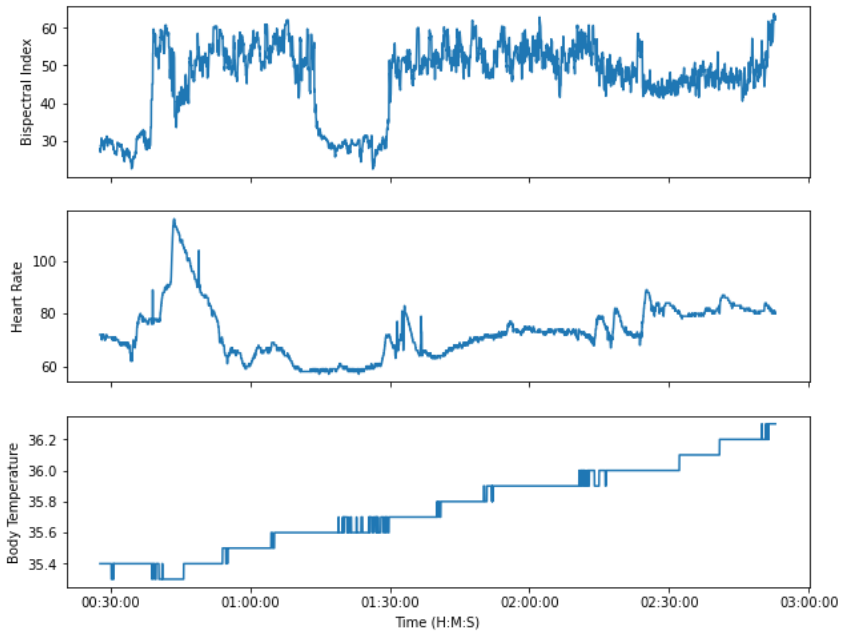

Figure 1: Patient 1 Biosignals during Surgery

Given that the inputs consisted of both discrete EHR data and sequential biosignal data, a model was developed that only fed EHR data through dense layers while the biosignal data was subject to LSTM layers as well. Figure 2 shows a visual representation of the model. The model was evaluated using a 80%-20% train-test split and hyperparameters such as the activation and objective functions were tuned. Finally, two outputs were evaluated: boolean ICU stay (i.e. whether or not a patient stayed in the ICU) and the length (in days) of the ICU duration.

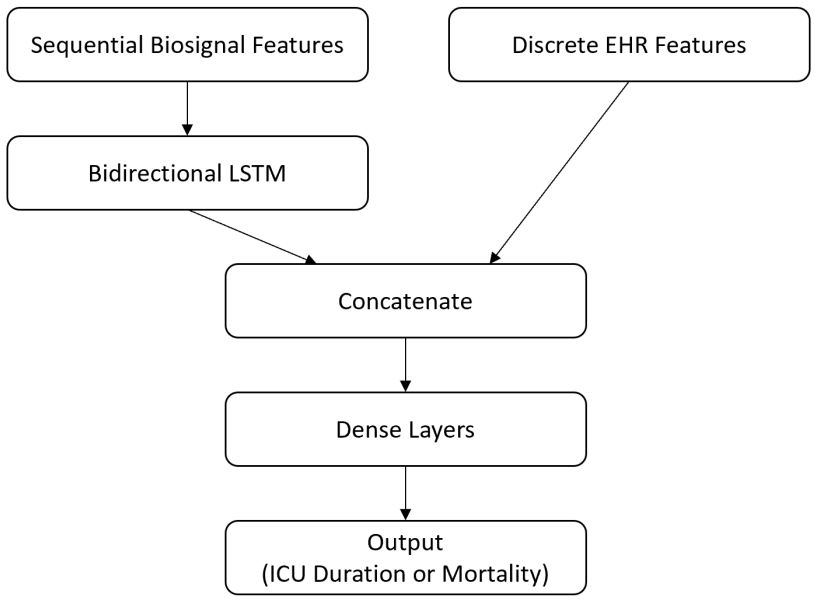

Figure 2: LSTM model architecture for ICU Duration and Mortality.

Mortality

Again, similar techniques were used to clean the EHR and biosignal data for this model as well. This model, however, included three major differences. First, class imbalance was a large issue for this model, with only 51 samples containing a positive class (i.e. the patient died). As such, synthetic minority oversampling technique (SMOTE) was implemented in order to synthesize new samples of the minority class (Appendix A). This increased the total number of samples to 9,544.

The other differences were related to the biosignal cleaning. After the biosignal data cleaning steps presented above were completed, two approaches were implemented to further reduce the size of the data. The first approach involved computing an exponential moving average (EMA) for each signal. The second approach involved frequency encoding each signal. In both cases, the results were a smaller input data size.

Results:

Results for each model are presented in the subsections below. A summary table of all Phase 2 results is presented within Table 5.

Patient Diagnosis

Using the methods described above, the Random Forest model achieved an accuracy of 46%, a 10% increase compared to Phase 1 results, with some classes achieving accuracies > 70%. Table 3 shows a sample of these results.

Table 3: Sample Patient Diagnosis Model Results

Operation Type	Diagnosis Accuracy	No. of Unique Diagnoses Present
Colorectal	25.73	34
Breast	31.25	5
Biliary/Pancreas	21.77	20
Stomach	74.35	4

In addition, a multi-output model was generated which displays the top three predicted diagnoses and their respective confidences for each case. This intent for this model is to help healthcare professionals make more informed decisions. Table 4 shows a sample of this model output. Finally, the important features used by each operation type were calculated, which could be used for further analysis and model improvements.

Table 4: Sample Predictions from Multi-Output Model

Sample	Prediction 1	Confidence 1	Prediction 2	Confidence 2	Prediction 3	Confidence 3
1	Hepatocellular carcinoma	22.4	Early gastric cancer	16.8	Varicose vein of lower limb	9.6

2	Advanced gastric cancer	15.8	Colon cancer	14.0	Rectal cancer	13.4
3	Early gastric cancer	15.4	Varicose vein of lower limb	12.0	Rectal cancer	11.4

It's important to note that these models only utilized EHR data as inputs, as opposed to biosignals, since the intent of this model is for hospitals to be able to use this prediction upon patient admission before any biosignals are collected.

ICU Length of Stay:

The LSTM model was tuned in order to predict both boolean ICU stay and ICU duration (in days). Ultimately, the model, as shown in Figure 2, utilized ReLu activations in the dense layers, and sigmoid activation in the final output layer. The binary cross entropy loss function was used for the boolean ICU stay response variable while the mean squared error loss function was implemented for the ICU duration response variable.. Several activation functions were evaluated, but showed minimal difference in performance.

After tuning the model, boolean ICU stay achieved an accuracy of 83% while ICU duration (in days) achieved an accuracy of 81%. Comparatively, the Phase 1 model results were 91% and 85%, respectively. Therefore, it was shown that the more complex model resulted in a decrease in accuracy.

Mortality:

After hyperparameter tuning, the LSTM model achieved an accuracy of 86% for the mortality response variable. Comparatively, during Phase 1, an XGBoost model was developed which achieved 95% accuracy.

In addition to the LSTM model architecture, the biosignal data was also used as inputs to an XGBoost model. As described in the Methods section above, after the biosignal data was cleaned, two techniques were used to further reduce the size of the data: frequency encoding and exponential moving averaging (EMA). Overall, the XGBoost model achieved 94% accuracy when using the EMA data as inputs and 96% accuracy when using the frequency encoded data. This represents a 10% improvement over Phase 1 results.

Table 5: Summary Phase 2 Results

Target Variable	**Model**	**Accuracy**
Patient Diagnosis	Random Forest	46%
ICU Stay (Boolean)	LSTM	83%
ICU Duration (days)	LSTM	81%
Mortality Rate	XGBoost	96%

Discussion:
The following subsections provide rationale for the results shown above as well as commentary on challenges and potential future work.

Patient Diagnosis:
Patient diagnosis, if predicted correctly, can help expedite the healing process of patients early. While model accuracy was improved from 36% in Phase 1 to 46% in Phase 2, unfortunately, this is still pretty low performance. However, the changes made during Phase 2, especially making the model a multi-output model, were positive changes that would hopefully give the healthcare provider more confidence in their decisions. In order to further improve the model, grouping the diagnoses in different ways could continue to be explored, especially if the algorithm designer has more domain knowledge. In addition, the feature importances tables could be used to further improve upon the model.

ICU Length of Stay:
Once a patient is admitted to the hospital, predicting boolean ICU stay or ICU duration (in days) is an important metric in order to help allocate and plan hospital resources. For boolean ICU stay, the best model utilized only EHR data and an XGBoost model architecture, resulting in an accuracy of 91%. In contrast, the biosignal data and LSTM architecture only achieved 83% accuracy. Similarly, the best model for ICU duration (in days) was observed during Phase 1 using EHR data and a Random Forest architecture which achieved 85% accuracy.

Given that the XGBoost and Random Forest models require less data, are less computationally expensive, and achieve higher accuracy, it's recommended to go forward with these Phase 1 models. If additional analysis into the LSTM model was desired, possible next steps could include additional cleaning and filtering of the biosignal data.

Mortality:
Mortality prediction can help hospital professionals better understand the current life status of patients, so as to allocate the scarce and important resources of the hospital reasonably. Overall, the best performing model utilized the processed biosignal data (i.e. frequency encoding) as inputs, an XGBoost model architecture and achieved an accuracy of 96%. Therefore, it was shown that the biosignal data was important to improving the model, however, the LSTM model architecture was unnecessary. In addition, it was shown that additional processing of the biosignal data, in the form of frequency encoding, was a positive addition to the pipeline.

Statement of Contributions:

- Parv Thakkar: Patient Diagnosis Model
- Tegan Ayers: Anesthesia Duration Model, Biosignal Data Cleaning, Initial LSTM Model
- Mehdi Safari: ICU Length of Stay Model
- Hongyang (Jarvis) Wang: Mortality Rate Model

References:

1. Lee HC, Park Y, Yoon SB, Yang SM, Park D, Jung CW. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Sci Data. 2022 Jun 8;9(1):279. doi: 10.1038/s41597-022-01411-5. PMID: 35676300; PMCID: PMC9178032.

Appendix A: SMOTE Oversampling Algorithm

(1) For each sample x in the minority class, using the Euclidean distance as the standard, calculate its distance from all samples in the minority class sample set to obtain its k-nearest neighbors.

(2) Set a sampling rate according to the sample imbalance rate, and determine the sampling multiple N. For each minority class sample x, randomly select several samples from its k-nearest neighbors, assuming that the selected neighbors are xn.

(3) For each randomly selected neighbor xn, use the original sample to construct a new sample according to the formula.

Formula: $x_{new} = x + rand(0,1) \times (\bar{x} - x)$

Appendix B: Exponential Moving Average (EMA)

Exponential moving averaging can be used to estimate the local mean of the variable, so that the update of the variable is related to the historical value within a period of time. The variable v is recorded as v_t at time t, and θ_t is the value of variable v at time t, that is, when the moving average model is not used $v_t = \theta_t$, after using the moving average model, the v_t update formula is as follows:

Formula: $v_t = \beta \times v_{t-1} + (1 - \beta) \times \theta_t$