Using blood glucose data to improved predictions of mortality for sepsis patients in the intensive care unit (ICU)

COMP90089: Machine Learning Applications for Health.
Final Report - Group 1
Semester 2, 2023.
University of Melbourne.

Hugo Lyons Keenan - 1081696
Sera Singha Roy - 1417176
Fathima Risla Jabarullah - 1138005
Tanvesh Sunil Takawale - 1230782
Arya Araban - 1439683

Abstract

This report investigates the usefulness of blood glucose levels and their variability in predicting mortality outcome of patients with sepsis, a critical medical condition with global significance. We review the relevant existing literature and formulate a hypothesis about the relevance of blood glucose to sepsis mortality. The study utilizes the MIMIC-IV database, employing various machine learning models within a 24-hour window for patient triage. The findings confirm the relevance of blood glucose levels in predicting sepsis patient outcomes. However, the trade-off between model complexity and interpretability is noted, and challenges in implementing these models in clinical settings are discussed. This project emphasizes the potential importance of blood glucose metrics in sepsis patient prognosis and the need to balance predictive power and interpretability in healthcare applications.

1 Introduction

Sepsis is a critical medical condition triggered by a systemic response to infection, potentially leading to organ dysfunction and failure. It is a significant global health issue, affecting millions annually and is a primary cause of ICU mortality worldwide [1]. Despite sepsis being a prominent cause of death in hospitalized patients, there is a limited amount of research on the predictive factors of mortality [2].

The primary objective of our project is to investigate the relationship between blood glucose levels and glucose variability on mortality in sepsis patients admitted to the ICU. Prompting this objective is a study by Rhodes et al. [3] that reported an increase in hyperglycemia among septic shock patients. This points to a relationship between glucose levels and sepsis that we may be able to exploit to create a better predictive model of the illness. It's important to emphasize that while the relationship between hyperglycemia and sepsis is of interest, the existing literature lacks documentation of a clear and definitive association.

Our focus is specifically on developing a model to assist with triage that may be used within the first 24 hours of admission to identify those at the highest risk of mortality [4]. This 24-hour window is a critical timeframe for assessing the influence of glucose levels and variability on patient outcomes.

2 Methods

2.1 Database used

We used MIMIC-IV database [5], a public healthcare records database containing ICU patient information taken from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Access to patient data is obtained in a legal and ethical manner, following the necessary data use agreements and obtaining appropriate approvals from relevant institutional review boards.

2.2 Digital Phenotyping

To create our patient cohort, we joined the `icustay_detail` table with the `mimiciv_derived.sepsis3` table using the `stay_id`, narrowing the dataset to those with sepsis. The `mimiciv_derived.sepsis3` table was created following the criteria set out in [6], and implemented for the MIMIC-IV dataset by the authors of [7]. We further narrowed our cohort to only include first ICU stays, patients aged 18 and above and a minimum ICU stay duration of 24 hours [4]. Furthermore, the query combined data from various tables (Figure 1) using appropriate keys like `stay_id`, `hadm_id`, and `subject_id` for effective organisation. Additionally, it employed data cleaning methods in computing important metrics like WBC count, glucose levels, and glucose variability, ensuring extreme values were properly handled, thereby improving data reliability (see 2.4). The query selectively chose attributes that are relevant to the research question, such as patient identifiers, demographic data (like gender and age), clinical scores (such as Charlson comorbidity index, APACHE III, SOFA), and clinical metrics (including WBC count, BMI, and glucose-related metrics). SOFA or Sepsis-related Organ Failure Assessment score [8] measures the degree of organ dysfunction of patients in the ICU over time; the time frame in MIMIC

is one hour. Apache III or apsiii [9] is a scoring system used in critical care medicine to assess the severity of illness and predict patient outcomes based on various physiological and clinical parameters. Values for these features are also taken from the MIMIC derived tables attributable to the authors of [7].

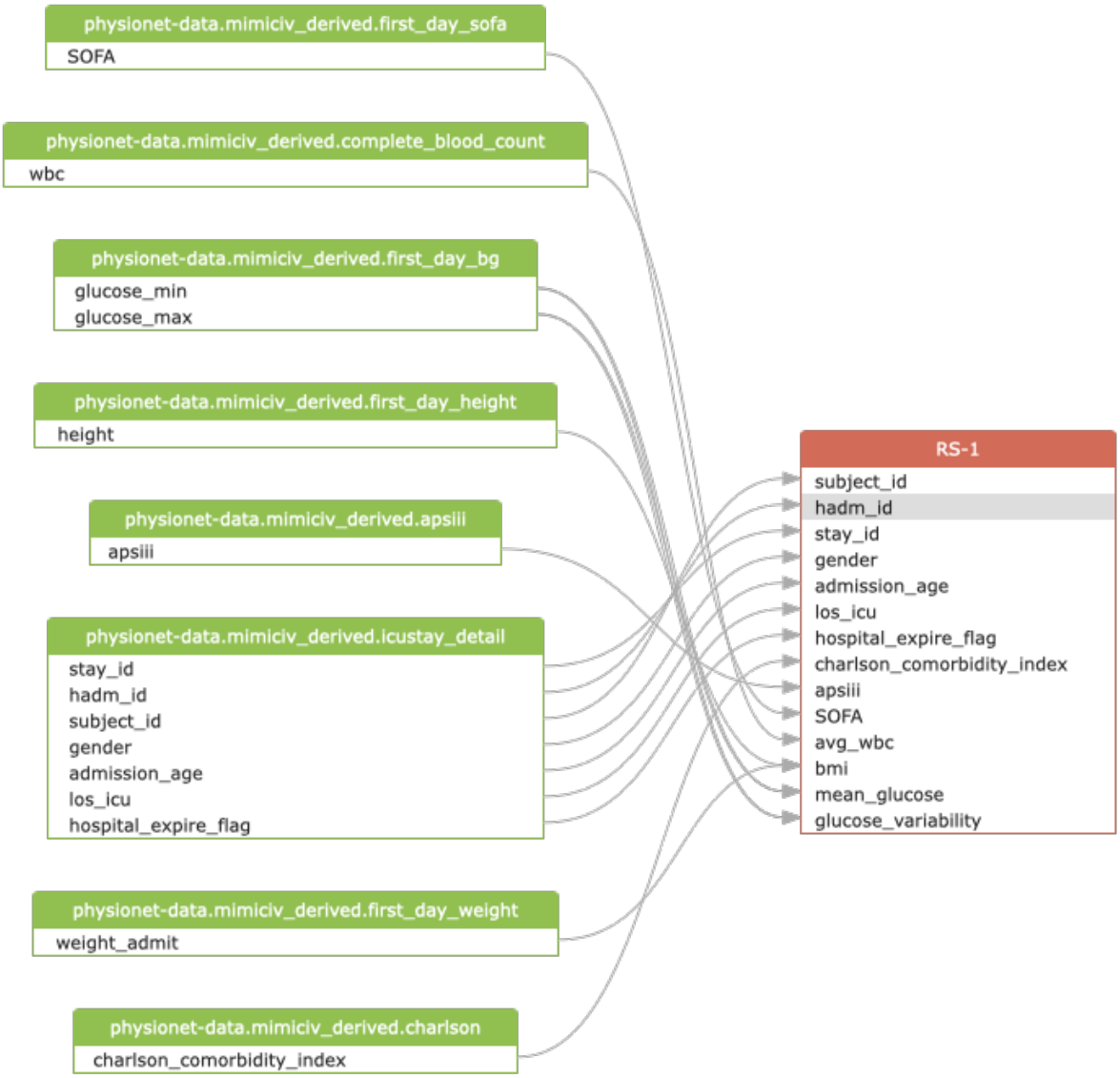

Figure 1: Digital Phenotype for the cohort.

2.3 Exploratory Data Analysis

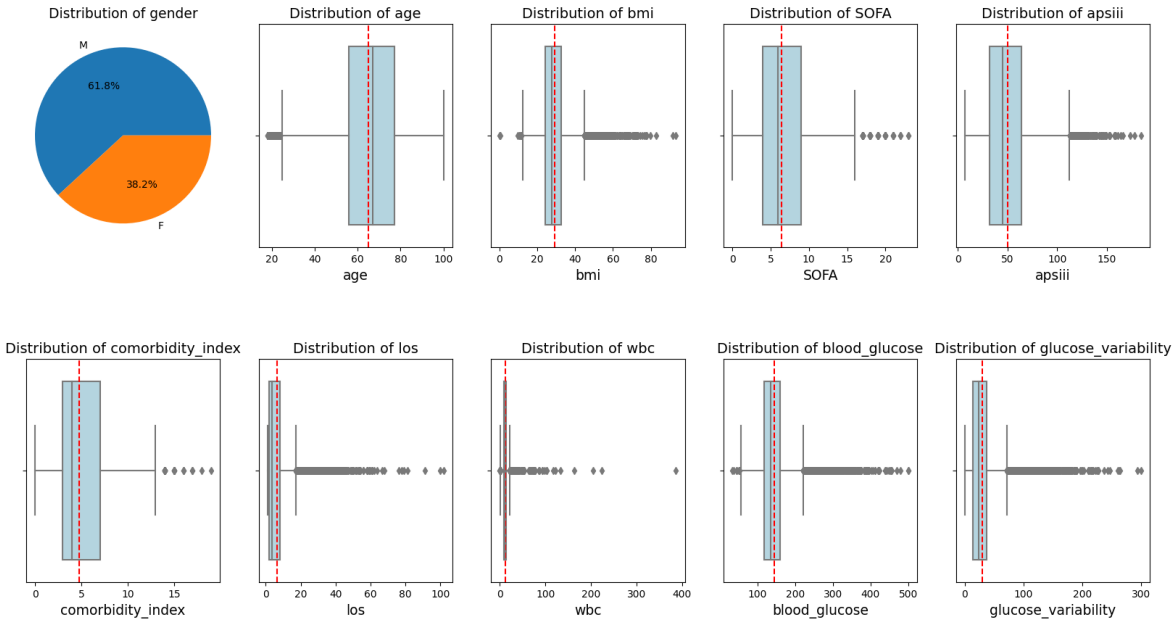

Figure 2: Cohort Feature Distribution

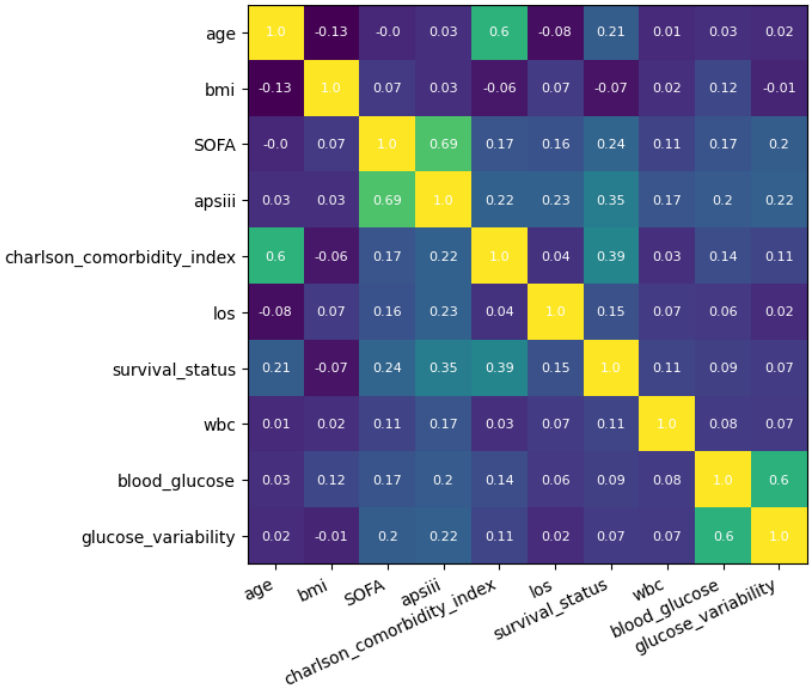

Figure 3: Cohort Feature Correlation

In this section we conduct exploratory analysis of our cohort. Figure 2 provides statistical information regarding the features, with the red lines representing the mean value of each feature. Figure 3 reveals relationships between various features and the target on a correlation heatmap. Some key insights are:

- **Survival Status**: Approximately 35.1% of the patients in our cohort who had sepsis did not survive.

- **Demographics**: The average age of the patients is around 65 years, suggesting an older population. Gender distribution is skewed, with males representing 61.8% of the cohort.
- **Body Mass Index (BMI)**: The average BMI is approximately 29.1, indicating that patients are generally overweight, but there is significant variability with many outliers.
- **Clinical Scores**: The SOFA and APSIII scores exhibit considerable variation, reflecting the diverse health conditions of ICU patients. The Charlson Comorbidity Index, which measures comorbid conditions, has an average value of about 4.7, indicating that patients typically have multiple pre-existing health issues.
- **Feature Correlations**: The correlation heatmap reveals strong positive correlations between age and Charlson Comorbidity Index, blood glucose levels, and glucose variability, as well as between SOFA and APS III scores. Additionally, patient survival status has moderate positive correlations with age, SOFA, APS III, and the Charlson Comorbidity Index, indicating that higher scores on these health indices are associated with higher mortality rates.

2.4 Preprocessing

The preprocessing steps employed in the dataset played a crucial role in enhancing data quality and ensuring the reliability of computed metrics. Extreme outliers in glucose values are addressed by capping them at 500, thereby preventing skewed statistics. Similarly, in the computation of BMI, an upper weight limit is imposed to 300 kg to mitigate the impact of extremely high weights. The 'gender' column is transformed into a boolean value, creating a new column named 'male.' Finally, data cleaning is applied to remove all rows with missing (NaN) values. Our final dataset is of size 11625. Additionally, for neural network training, standardisation is performed to conform all features with a mean of 0 and a standard deviation of 1. We also employ techniques to handle the class imbalance (see section 4.2)

2.5 Predicitive Models

In our attempts to predict the mortality rate of sepsis patients in the ICU, a number of predictive models were experimented with. Note that we do not use length-of-stay (LoS) as a feature in these models as it is not available at the time they will be used. The models we consider include:

- **Logistic Regression:** A straightforward model offering interpretability through direct feature weighting. Often struggles with complex, non-linear patterns.
- **Decision Tree:** Rule-based and interpretable, provide a simple measure of feature importance but may lack complexity in modelling.
- **Random Forest:** An ensemble of decision trees that enhances prediction and generalization, reducing overfitting and providing feature importance metrics.
- **Support Vector Machine (SVM):** Utilizes linear and non-linear kernels to separate classes with a hyperplane in a high-dimensional space.
- **Gradient Boosting:** An ensemble approach like Random Forest, but iteratively focuses on challenging cases to improve model performance.
- **Neural Network (NN):** Highly expressive, capable of modeling intricate relationships but potentially prone to overfitting and usually less interpretable.

Each of these models were evaluated using accuracy, precision and recall (for each class). The objective was to select the model that would be maximally helpful for clinicians in a critical care setting.

3 Results

Summary results for each model are described in table 1. These results are covered in more detail in section 4. We also explain two of our models in detail below.

Classifier	Accuracy	Precision		Recall	
		Recovered	Died	Recovered	Died
Decision Tree	64.00%	73%	48%	71%	50%
Logistic Regression	71.53%	83%	57%	72%	71%
Random Forest	74.62%	79%	65%	84%	57%
SVM	71.56%	82%	57%	72%	70%
Gradient Boosting	72.55%	83%	58%	73%	73%
Neural Network	72.93%	83%	59%	73%	73%
NN : $w_p = 1$	**75.41%**	77%	**66%**	**85%**	56%
NN : $w_p = 9.27$	65.72%	**88%**	51%	55%	**86%**

Table 1: Metrics for various classifiers. Unless otherwise stated, balanced class weights are used.

3.1 Logistic Regression

Our logistic regression model is one of our most basic and interpretable models providing a baseline to which we can compare other models. We can see in table 2 the coefficient estimates of each of our features, as well as their corresponding z scores and implied significance. Unsurprisingly, each of SOFA, APSIII and CCI have statistically significant, positive coefficients, meaning higher values will lead to higher likelihoods of death. We see that blood glucose also has a positive coefficient which is statistically significant, meaning higher readings are also correlated with lower odds of survival. Interestingly, we see that glucose variability has a negative coefficient, suggesting that those with greater variation in glucose levels have better odds of survival. Note however that this is not a statistically significant result and thus should not be read deeply into.

Coefficient	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	−4.883	0.224	−21.845	< 2e-16 ***
age	−0.002	0.002	−0.773	0.439466
bmi	−0.015	0.004	−3.834	0.000126 ***
SOFA	0.062	0.010	6.001	1.96e-09 ***
apsiii	0.027	0.002	16.538	< 2e-16 ***
charlson_comorbidity_index	0.154	0.012	12.960	< 2e-16 ***
wbc	0.063	0.005	13.208	< 2e-16 ***
blood_glucose	0.002	0.001	2.950	0.003174 **
glucose_variability	−0.002	0.001	−1.529	0.126226
male (True)	−0.224	0.059	−3.817	0.000135 ***

Table 2: Coefficient estimates with standard errors, z-values, and p-values.

3.2 Neural Network

We also chose to implement a basic neural network, a multilayered perceptron, in an attempt to capture non-linear patterns in the data and improve accuracy. The basic model has 2 fully connected hidden layers of size 100 and 50, with a single output neuron that may be read as the positive class probability. Each fully connected layer is passed through a RELU function and the output neuron is passed through a sigmoid function to normalise its value to $[0, 1]$.
Our neural network is our best performing model in terms of accuracy however this may not be the most appropriate metric, see section 4 for discussion.

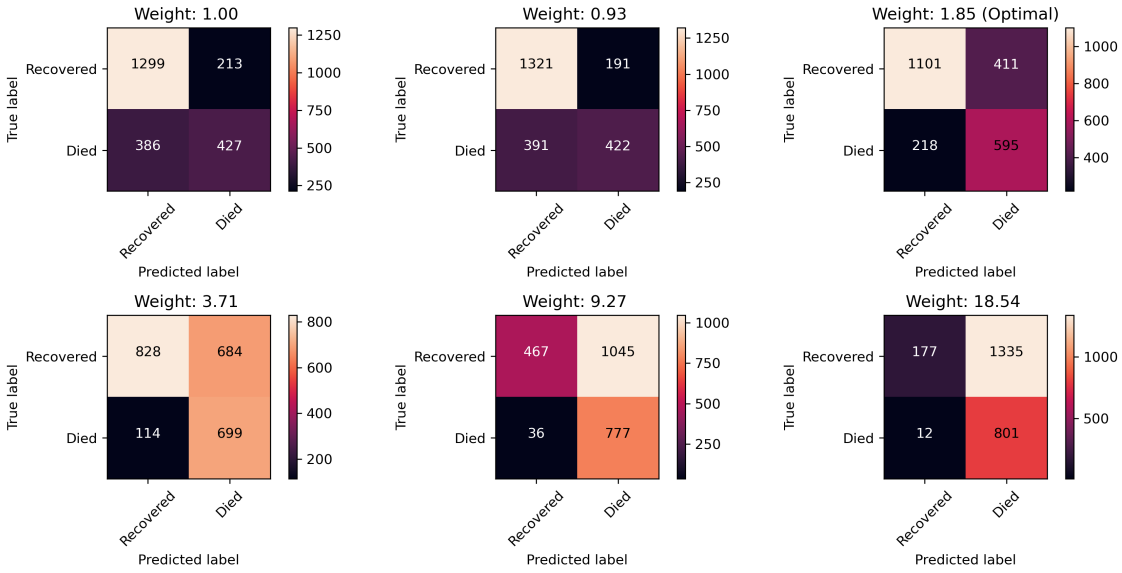

Figure 4: Confusion matrices for different positive class weights for NN.

4 Discussion

In this section we give a brief discussion of our experiments along with some key findings.

4.1 Explainability/Simplicity vs. Predictive Power

In our experiments, we found that the more complex, less interpretable models, such as random forests and neural networks, tended to give better performance according to various metrics that we tested them on. When choosing which models to roll out in practice however, we may be inclined to choose a simpler model even if it has slightly lower accuracy. This has the advantage of being both more interpretable to the clinicians making the decisions and more robust when encountering data that may be out of the training distribution, increasing stability and trust in the model.

4.2 Class Balance

A notable difficulty of this project was dealing with the significant class imbalance in the dataset. As mentioned in section 2.3, more examples exist of individuals who survive their stay (%64.9) than not, incentivising our model to predict the majority class more of the time. To combat this, we modify our training process to assign a higher weight to positive examples (deaths). In the case of neural networks, we modify the standard cross entropy loss function to have a weight ($\neq 0$) associated with the positive class[1]:

$$L(y, \hat{y}) = -(1-y)\log(1-\hat{y}) - w_p y \log(\hat{y}) \tag{1}$$

where:

- $L(y, \hat{y})$ is the loss function comparing the true label y and the predicted probability $\hat{y}$.
- $y \in \{0, 1\}$ is the true label, indicating the class of the sample.
- $\hat{y} \in [0, 1]$ is the predicted probability, outputted by the model, indicating the likelihood of the sample being in the positive class.
- w_p is the weight for the positive class.

[1]This is usually set to the inverse proportion of positive labels in the dataset

The total loss over a batch of N examples is usually the mean of the losses for individual examples:

$$L_{\text{batch}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{y}_i) \tag{2}$$

We can observe the results of this for different weightings in figure 4, where we can see that higher weights for the positive class cause the classifier to strongly avoid predicting recovery for any patient, as false negatives are increasingly punished in the loss function.
We see similar results when adjusting the class weights for other models, where more weighting on the minority class leads to it's recall increasing at the cost of precision. See section 4.4 for a discussion of how this relates to the hospital context.

4.3 Random Forest Feature Importances

After training our random forest classifier, we can look at the model's internally calculated importance for each feature to gauge the role being played by an individual's glucose level and variability. In figure 5, we can see that mean glucose levels, as well as glucose variability, play a meaningful role in our final classifier. This suggests that despite no strong first-order effect being discovered in the logistic regression coefficients (Table 2), glucose data may play an essential role in differentiating sub-cohorts that have been split based on other features.

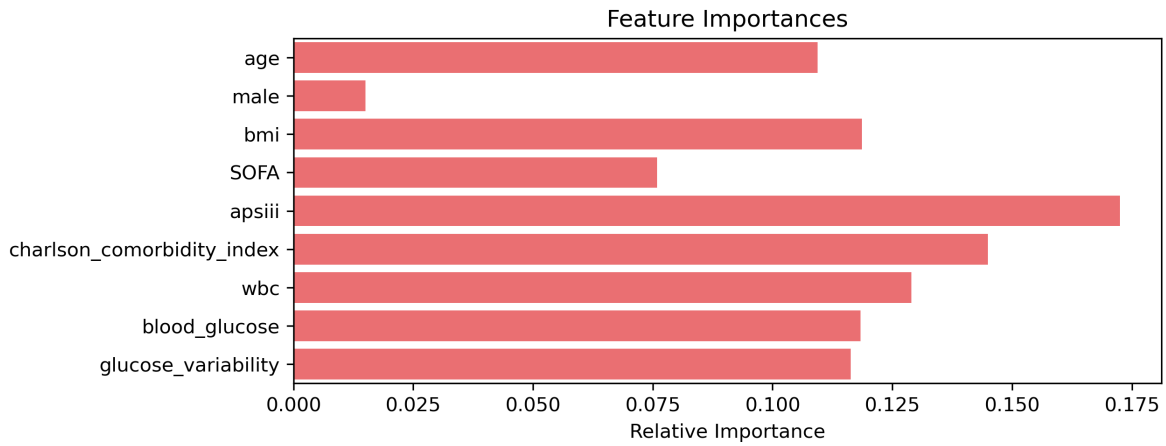

Figure 5: Feature Importances for Random Forest Classifier.

4.4 Limitations and Feasibility

Like many applications of ML in healthcare, applying this project's predictive models in a clinical setting comes with many challenges. In healthcare, particularly in intensive care units (ICUs), the significance of false negatives cannot be overstated. Our optimal neural network model has a substantial number of false negatives (218) indicating the risk of failing to identify critical patients. This could delay vital interventions, potentially leading to worsened outcomes or even death. Additional barriers include privacy concerns, rigorous clinical validation, integration with electronic health records, gaining clinician acceptance and lack of model interpretability. Additionally, regulatory approvals and resource constraints need to be addressed to implementation. Overcoming these barriers will require collaborative efforts between data scientists, clinicians, and healthcare institutions to ensure the successful and ethical use of the predictive models in real-world healthcare environments.

5 Conclusion

This study investigated the relationship between blood glucose levels, glucose variability, and mortality in ICU sepsis patients. The findings revealed significant patterns and correlations, suggesting the potential for their use in predicting patient outcomes. The analysis considered various factors, such

as age, co-morbid conditions, organ failure scores, and blood glucose metrics. Notably, higher blood glucose levels and increased variability were found to be useful in predicting mortality.
Different predictive models were employed, with more complex models exhibiting higher accuracy. However, a trade-off between model complexity and interpretability was noted. The study highlighted the challenge of class imbalance in the dataset and the importance of handling it effectively.

In summary, this project underscores the potential importance of blood glucose metrics in predicting sepsis patient mortality. It emphasizes balancing predictive power with interpretability in real-world clinical applications. Future work could include improving the model through various model optimization methods and ensemble methods to achieve higher accuracy and lower false negatives.

6 Code

The code, the data used and a README file is available here.

7 Contributions

Team Member	**Contribution**
Hugo Lyons Keenan	Methodology, Project administration, Software, Writing – original draft, Writing – review & editing
Sera Singha Roy	Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing
Fathima Risla Jabarullah	Investigation, Formal Analysis,, Methodology, Software, Writing – original draft, Writing – review & editing
Tanvesh Sunil Takawale	Investigation, Formal Analysis, Resources, Writing – review & editing
Arya Araban	Methodology, Software, Visualization, Writing – review & editing

Table 3: Team Member Contributions

References

[1] Fan SL, Miller NS, Lee J, Remick DG. Diagnosing sepsis – The role of laboratory medicine. Clinica Chimica Acta. 2016;460:203-10. Available from: https://www.sciencedirect.com/science/article/pii/S0009898116302935.

[2] Mohan A, Shrestha P, Guleria R, Pandey R, Wig N. Development of a mortality prediction formula due to sepsis/severe sepsis in a medical intensive care unit. Lung India : official organ of Indian Chest Society. 2015 07;32:313-9.

[3] Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. Intensive Care Medicine. 2017 Jan;43(3):304-77. Available from: https://doi.org/10.1007/s00134-017-4683-6.

[4] Ali NA, O'Brien JM Jr, Dungan K, Phillips G, Marsh CB, Lemeshow S, et al. Glucose variability and mortality in patients with sepsis. Crit Care Med. 2008 Aug;36(8):2316-21.

[5] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV. PhysioNet; 2021. Available from: https://physionet.org/content/mimiciv/1.0/.

[6] Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, et al. Developing a New Definition and Assessing New Clinical Criteria for Septic Shock: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016 02;315(8):775-87. Available from: https://doi.org/10.1001/jama.2016.0289.

[7] Johnson AEW, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. Journal of the American Medical Informatics Association. 2017 09;25(1):32-9. Available from: https://doi.org/10.1093/jamia/ocx084.

[8] Vincent JL, Moreno RPJ, Takala J, Willatts SM, de Mendonça A, Bruining HA, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine. 1996;22:707-10. Available from: https://api.semanticscholar.org/CorpusID:40396839.

[9] The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults - PubMed — pubmed.ncbi.nlm.nih.gov;. [Accessed 02-11-2023]. https://pubmed.ncbi.nlm.nih.gov/1959406/.