Prediction and Data Visualization of Sepsis

Yantao Luo
New York University
New York, NY
yl5929@nyu.edu

Yihan Li
New York University
New York, NY
yl10798@nyu.edu

ABSTRACT

Sepsis is a critical condition that poses a complex challenge to healthcare systems, having severe implications for patient outcomes and resource allocation. This paper presents an innovative approach to the early prediction of sepsis, utilizing advanced machine learning techniques to enhance the timely detection and treatment of this life-threatening condition. We developed and evaluated three predictive models: K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest, each demonstrating varying degrees of success in distinguishing sepsis cases from non-sepsis instances within a comprehensive clinical dataset.

Our study involved a rigorous methodology, including extensive exploratory data analysis, data preprocessing for normalization and outlier removal, and a two-stage data imputation strategy to address missing values prevalent in the dataset. We employed performance metrics such as accuracy, precision, recall, F1-score, and others, to provide a comprehensive evaluation of the models' predictive capabilities.

The results indicate that each model holds potential for clinical application, with the Random Forest model showing particular promise due to its robust performance across various metrics. We discussed the implications of these findings, emphasizing the necessity for continuous model refinement, exploration of cost-sensitive learning, and the integration of predictive analytics into clinical workflows. We conclude that the application of such machine learning models can substantially mitigate the adverse outcomes of sepsis and potentially other diseases, marking a significant advancement towards data-driven healthcare solutions.

1 INTRODUCTION

Sepsis is a severe medical condition characterized by a systemic inflammatory response to infection, which can lead to organ dysfunction and, in some cases, death [4]. It poses a significant threat to patients, with a mortality rate ranging from 25% to 30%, often resulting from complications such as hypotension, hypoxemia, metabolic acidosis, and more [11]. In 2017 alone, there were approximately 48.9 million reported cases of sepsis worldwide, leading to 11 million sepsis-related deaths [10].The paradox of sepsis lies in the body's own defense mechanism turning detrimental, where the infection-fighting process wreaks havoc, impairing organ function and, in severe cases, leading to septic shock. This precarious drop in blood pressure signifies a medical emergency that can inflict extensive damage to vital organs, including the lungs, kidneys, and liver, further complicating recovery prospects.

Given the critical threats of Sepsis, timely sepsis prediction can play a pivotal role in preserving organ function and increasing survival rates. However, the prediction of early sepsis is a multifaceted challenge. On the one hand, false positive prediction in non-septic patients or exceedingly early detection in septic patients can strain limited hospital resources. On the other, delayed recognition of sepsis can accelerate the patient's decline towards a life-threatening state, drastically reducing the chances of a positive outcome.

This paper explores the development and application of advanced machine learning models to address the urgent need for early and reliable sepsis detection. By leveraging predictive analytics and data visualization techniques, we aim to offer a potential solution that could transform patient care and outcomes. By employing a dataset rich with clinical parameters, we have trained three predictive models: K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest. The performance of each model has been thoroughly evaluated for its ability to distinguish between positive and negative sepsis cases, providing insights into their potential integration into clinical practice.

Therefore, the main contributions of this project are:

- **Data Visualization Strategies for Clinical Insights:** We leverage various data visualization techniques to extract and communicate key clinical insights from the data, aiding in the interpretation and understanding of sepsis indicators.
- **Predictive Models for Early Sepsis Detection:** Our models utilize a range of statistical and machine learning techniques to predict the onset of sepsis, aiming to provide healthcare professionals with a valuable predictive tool.
- **Publicly Available Code:** In the spirit of transparency, we have made our full code available on Github, allowing for replication, verification, and further exploration.

The following sections will outline the materials and methods employed in developing our predictive models, delve into the intricacies of the data visualization strategies utilized, and discuss the implications of our findings within the clinical context. Through this study, we aim to contribute a significant tool in the ongoing battle against sepsis.

2 METHODOLOGY

In this section, we present our methods of dataset analysis, model construction, and evaluation. All of the code is written in Python using Jupyter Notebook.

2.1 Dataset

The dataset selected for this project was retrieved from Kaggle [13]. It offers a comprehensive array of vital signs, laboratory values, demographics, and outcomes related to sepsis. It is specifically designed to challenge participants in predicting sepsis occurrences six hours before clinical prediction becomes possible. Additionally, the dataset is structured to reward early predictions, penalize late predictions, and discourage false alarms, aligning with the critical need for timely detection and intervention in sepsis cases.

- Vital Signs: These are the first-line indicators in the clinical assessment of a patient. The dataset includes eight distinct vital sign measurements such as Heart rate (HR), Oxygen saturation (O2Sat), Temperature (Temp), Systolic Blood Pressure (SBP), Mean Arterial Pressure (MAP), Diastolic Blood Pressure (DBP), Respiration rate (Resp), and End-tidal Carbon Dioxide (EtCO2).
- Laboratory Values: Reflecting the internal biochemical status of patients, the dataset provides a vast array of laboratory measurements. These indicators range from blood pH levels to the partial pressure of carbon dioxide (PaCO2).
- Demographics: Patient demographics are critical in assessing the risk and prevalence of sepsis across different population segments. The dataset includes age, gender, administrative identifiers for ICU units, time intervals between hospital admission and ICU transfers, as well as the length of stay in the ICU.
- Outcome: The SepsisLabel column serves as the ground truth for the presence or absence of sepsis, marked against a specific time threshold to aid in early prediction analysis.

2.2 Exploratory Data Analysis

2.2.1 Statistics of the dataset. The csv file contains 790,215 rows and 43 columns, including 41 columns of features, 1 column of patient id, and 1 column of label.

Records of 20,336 patients are collected. 58.2 % of the patients are male, and 41.8% of them are female (Figure 1). Most of the patients are aged from 40 to 90 (Figure 2). In average, each patient was observed for 38.86 hours. And 2.2 % of the records are recognized as Sepsis (Figure 3). We also plot the histograms of all the features and display some of those histograms (Figure 4).

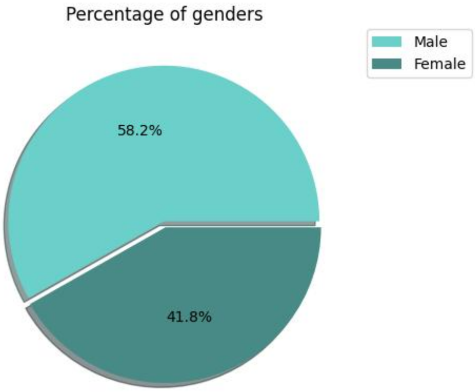

Figure 1: Percentage of genders

2.2.2 Nullity of the dataset. The dataset shows a rather high nullity. The average percentage of null values is 65.04 %. 20 columns have 90 % or more null records. We show a heatmap of missing data of each features (Figure 5). To address this problem, we performed data imputation before training our predictive models. (See the details in Section 2.3)

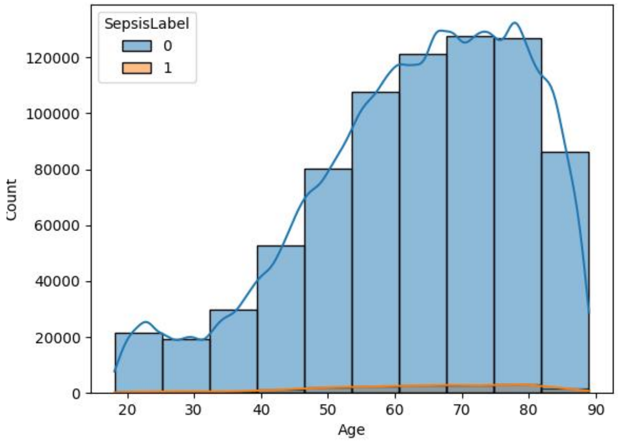

Figure 2: Histogram of Ages of Patients

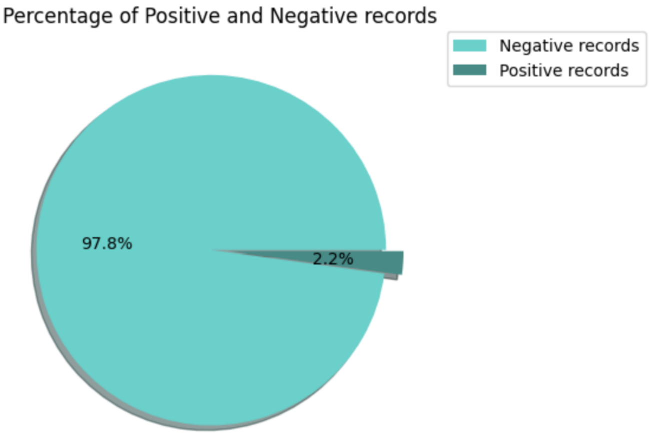

Figure 3: Percentage of positive Sepsis cases

2.2.3 Box plot Analysis. We plot the box plots of each clinical values. As shown in Figure 6, there are a large number of outliers in many features. We managed to remove those outlier in the preprocessing stage.

2.3 Data Preprocessing

2.3.1 Data scaling. In the process of preparing our dataset, we observed that some features contain values across a wide range. These features did not conform to a normal distribution, which is a common prerequisite for many statistical models and machine learning algorithms [2]. To address this problem, we applied a logarithmic transformation to those features. Later, we normalized all the features.

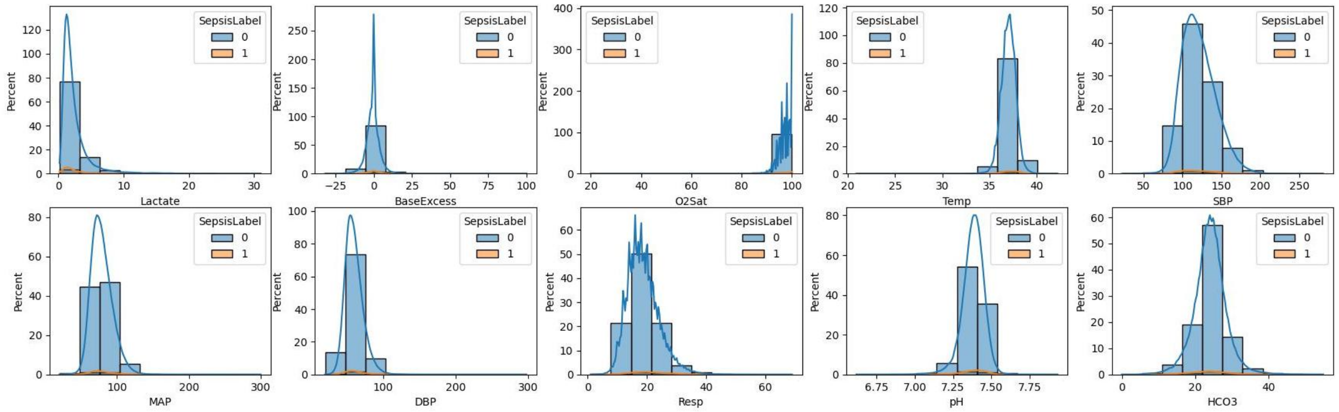

Figure 4: Histograms for part of features

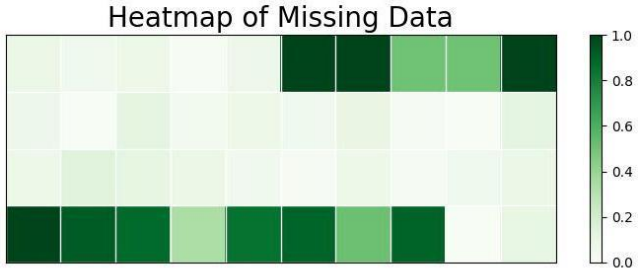

Figure 5: Heatmap of missing data (lighter color denotes more missing data)

2.3.2 Remove Outliers. We then removed outliers of each feature using a threshold of zscore equals to 3.

2.3.3 Data imputation. To address the challenge of missing values in such datasets, we employed a robust preprocessing technique using the bfill() and ffill() methods from the Pandas library. This approach is reasonable and scientifically sound for datasets derived from the continuous monitoring of patients. It is important to note that all imputations must be performed within the groups of each individual patient.

The imputed data still contained null records, due to the overwhelming amount of missing data. Consequently, we dropped the features with ratios of null records higher than 50%, and then further imputed the dataset with mean values on a patient-by-patient basis.

2.3.4 Split of training set and test set. We adopted a random split with a ratio of 0.8 versus 0.2.

2.3.5 Resample the imbalanced dataset. Our dataset was significantly imbalanced, with only 2.2% of records being positive cases of Sepsis. Such a severe imbalance can lead to biased model learning [7]. To address this issue, we employed the SMOTEENN resampling technique to achieve a balanced dataset [1]. This resampling resulted in approximately a 1:1 ratio between the majority and minority classes, thereby mitigating the risk of bias in our model's training process.

2.3.6 Feature selection. We presented a heatmap of correlation matrix between the independent variable *SepsisLabel* against all features. As shown in Figure 7, no features exhibit significant correlation with *SepsisLabel*. Hence, we dicided to keep all the features.

2.4 Classifiers

In order to accurately predict sepsis, we employ three classifier models, each contributing its unique strengths to our predictive analysis. These models have been selected for their proven efficacy in health informatics and their ability to handle complex datasets, such as the one we used for this sepsis prediction. The following classifier models have been adopted for this project:

- K-Nearest Neighbors (KNN): The KNN algorithm is a non-parametric method used for classification and regression. By measuring the distance between data points, KNN classifies a data point based on how its neighbors are classified. This method is intuitively simple yet powerful, allowing for an adaptable model that aligns with the dynamic nature of clinical data. Its efficacy in detecting patterns in similar patient profiles makes it an invaluable tool in the early detection of sepsis, especially when considering the nuanced variations in patient vital signs and lab results [12].
- Logistic Regression: This model is particularly advantageous for binary classification problems [3]. Logistic regression estimates the probability of a binary outcome based on one or more predictor variables, making it a staple in the field of medical statistics. It is especially useful for understanding the relationship between the binary characteristic of sepsis onset and continuous and categorical predictor variables, such as those present in our dataset. By providing probabilities along with binary outcomes, it offers a clear framework for decision-making in clinical settings.
- Random Forest: As a robust ensemble learning method, the Random Forest algorithm constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. This method

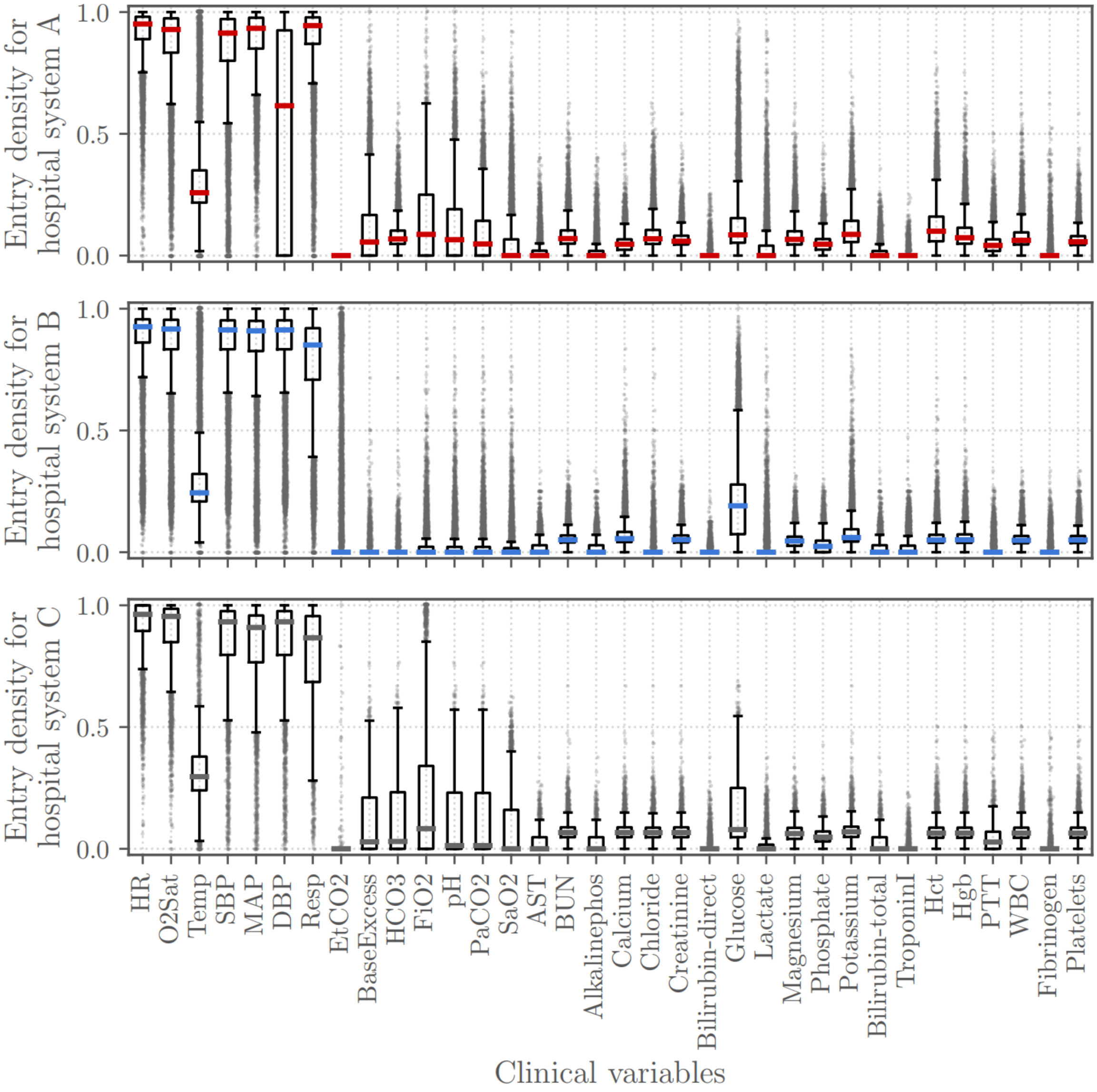

Figure 6: Box plots for each feature (Image source: [9])

is known for its high accuracy, capability to handle large datasets with higher dimensionality, and its power to model complex interactions and classifications. It is also widely applied in health care classification tasks [5] [6].

Each of these classifiers will be trained and validated on a comprehensive set of features extracted from vital signs and laboratory values, ensuring a rigorous and thorough analysis. All of the models are implemented using Scikit-learn (sklearn) library [8].

2.5 Performance Metrics

To comprehensively evaluate the performance of our predictive models, we employ a suite of metrics that provide insights into various aspects of model accuracy and generalizability. The chosen metrics are pivotal for interpreting the models' predictive capabilities, especially given the imbalanced nature of our dataset where the number of sepsis-positive cases is significantly lower than the number of negative cases. The following metrics are used:

- Accuracy: This metric provides an overall success rate of the model by measuring the proportion of true results, both true positives and true negatives, in the total dataset. It is a useful initial indicator of model performance, especially when dealing with relatively balanced datasets.
- Precision: Precision is critical in the context of medical predictions as it reflects the proportion of true positive identifications out of all positive identifications made by

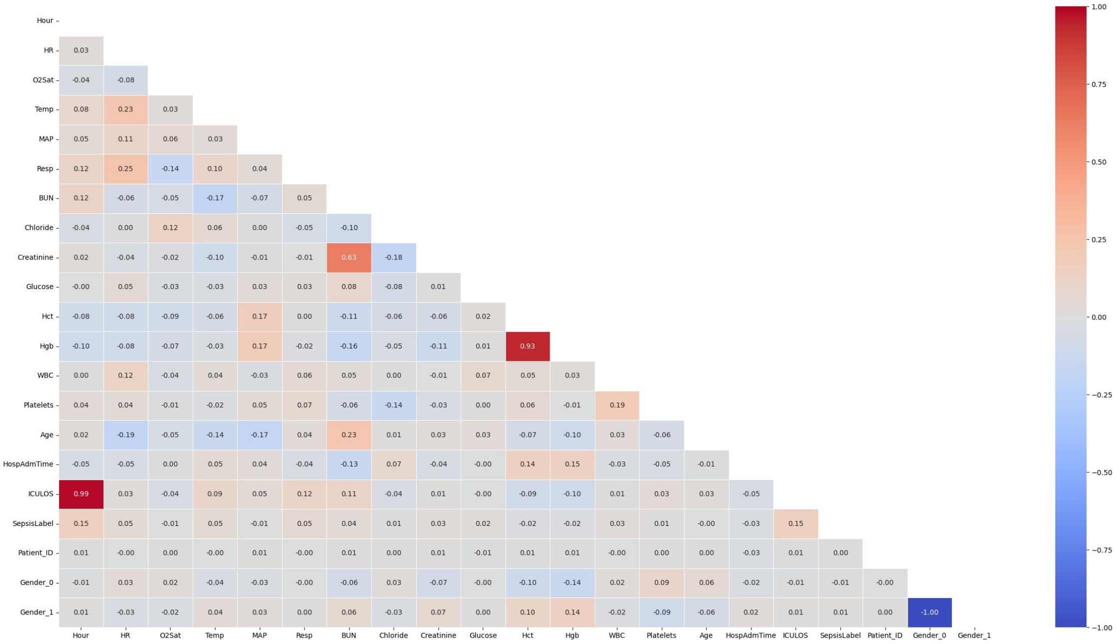

Figure 7: Heatmap of correlation matrix (the independent variable is *SepsisLabel*)

the model. High precision indicates a lower rate of false positives, which is crucial in avoiding unnecessary treatments in a clinical setting.

- Recall: Recall assesses the model's ability to correctly identify all actual positives. In the domain of sepsis prediction, a high recall rate is desirable to ensure that the majority of sepsis cases are identified, allowing for timely treatments.
- F1-Score: Given that precision and recall are often in a trade-off, the F1-score is utilized to find the harmonic mean between these two metrics. It is particularly valuable when seeking a balance between the model's precision and recall, providing a single score that gauges the model's accuracy in identifying the positive class while considering the incidence of false positives and false negatives [14].
- Support: This metric indicates the number of actual occurrences of each class in the specified dataset, which is used to contextualize the evaluation and provide insight into the class distribution that the model was exposed to during its training.
- Macro Average: The macro average calculates the arithmetic average of the metric for each class, treating all classes equally. This average is unaffected by class imbalance, thus providing an assessment that does not overemphasize the majority class.
- Weighted Average: The weighted average takes into account the imbalance by weighting the metric score of each class by the number of true instances of each class. This gives a more accurate reflection of the model's performance across the dataset as it takes class distribution into consideration.

These metrics collectively guide us in refining our models. By using multiple metrics for model evaluations, we ensure a well-rounded assessment of model performance, enabling us to identify strengths and weaknesses that may not be apparent through a singular metric.

3 RESULTS

3.1 K-Nearest Neighbors

The evaluation of the K-Nearest Neighbors (KNN) model shown in Table 1 demonstrates a high level of accuracy, suggesting that it performs well in distinguishing between sepsis and non-sepsis cases within the test dataset. The model shows a commendable ability to identify true sepsis cases, while also maintaining a substantial rate of correct predictions for sepsis. The F1-score indicates a robust predictive performance for both classes, albeit slightly lower for the positive class, reflecting the model's effectiveness in classifying cases as septic or non-septic. Overall, the model's metrics present a strong predictive capability, balanced by the weighted averages that

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	154639
1.0	0.70	0.90	0.79	3404
accuracy			0.99	158043
macro avg	0.85	0.94	0.89	158043
weighted avg	0.99	0.99	0.99	158043

Table 1: The classification report of our KNN model

	precision	recall	f1-score	support
0.0	0.99	0.72	0.83	154639
1.0	0.05	0.64	0.09	3404
accuracy			0.72	158043
macro avg	0.52	0.68	0.46	158043
weighted avg	0.97	0.72	0.82	158043

Table 2: The classification report of our Logistic Regression model

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	154639
1.0	0.88	0.87	0.88	3404
accuracy			0.99	158043
macro avg	0.94	0.93	0.94	158043
weighted avg	0.99	0.99	0.99	158043

Table 3: The classification report of our Random Forest model

account for class distribution, thus affirming its potential utility in a clinical setting for early sepsis detection.

3.2 Logistic Regression

The Logistic Regression model's performance, as summarized in Table 2, indicates a satisfactory level of accuracy, with nearly 71.88% of predictions aligning with the actual data. The precision and recall for the non-sepsis predictions are high, but the precision for the sepsis predictions (class 1) is notably lower, which indicates a significant number of non-sepsis cases are being incorrectly labeled as sepsis. Despite this, the model still retains a reasonably good recall for class 1, indicating that it can identify a majority of the true sepsis cases, albeit with some false alarms. The F1-score for the sepsis class is on the lower end, which could be a concern in a clinical setting where the cost of false positives can be high.

Therefore, while the Logistic Regression model is adept at recognizing non-sepsis cases, it struggles with accurately identifying sepsis cases without generating a substantial number of false positives. This indicates a need for more sophisticated or balanced approaches for handling the imbalanced class distribution inherent in the dataset.

3.3 Random Forest

The evaluation of the Random Forest model, as shown in Table 3, suggests strong performance across the board. With near-perfect precision and recall for the non-sepsis class 0 and high scores for the sepsis class 1, the model demonstrates a high degree of accuracy in classifying both conditions. The balanced F1-score for the sepsis class reflects an effective harmony between precision and recall, indicating the model's capability in correctly identifying sepsis cases as well as its precision in minimizing false positives.

The accuracy of the model stands at 94%, a robust indicator of overall performance, while the macro and weighted averages for precision, recall, and F1-score are all above 93%. These averages suggest that the model's predictions are consistent and reliable across different classes, despite the potential imbalance in the dataset. Hence, the model appears to be an effective tool for sepsis prediction, potentially being useful in a clinical setting for early identification of sepsis.

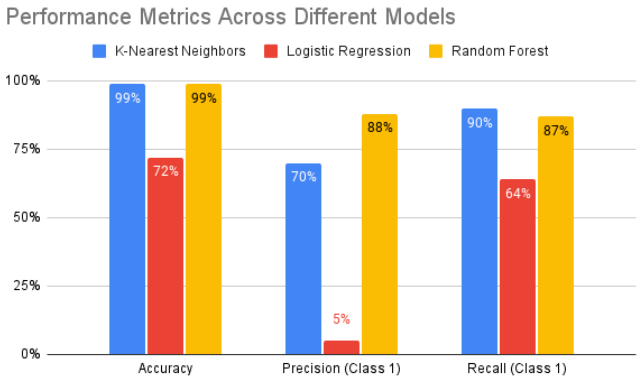

Figure 8: Performance Metrics Across Different Models

4 DISCUSSION

The implementation of machine learning models in the clinical setting for the prediction of sepsis represents a significant advancement in patient care. Our models—K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest—have each demonstrated a capacity to distinguish between sepsis and non-sepsis cases with varying degrees of success, as shown in Figure 8. Their potential application in clinical environments could aid in the early detection of sepsis, allowing for timely interventions that could save lives and reduce healthcare costs associated with late sepsis management.

The KNN model's high accuracy and balanced performance across precision and recall suggest that it could serve as a reliable tool for initial screening. The Logistic Regression model, while displaying satisfactory accuracy, highlights the complexity of sepsis prediction due to its lower precision for sepsis cases, indicating a need for further model refinement to reduce false positives. The Random Forest model's strong performance across all metrics, including a high F1-score for the sepsis class, indicates its robustness and potential as a frontline predictive tool in a clinical dashboard.

To enhance the predictive performance of our models, several strategies could be considered. The incorporation of more granular clinical data, such as patient medical history and treatment responses, could provide a richer dataset for model training. Advanced feature engineering techniques and the exploration of non-linear relationships within the data may yield more nuanced insights.

Moreover, the application of ensemble methods could leverage the strengths of each individual model, potentially leading to better generalization and performance.

The models could also be improved by implementing cost-sensitive learning approaches to address the imbalance in the dataset. Given the higher cost of false negatives in sepsis prediction, adjusting the models to emphasize sensitivity over specificity could be beneficial. Additionally, continuous model validation using up-to-date and diverse datasets from various healthcare settings will be crucial to ensure that the models remain relevant and effective across different patient populations.

In a clinical context, the interpretation of model predictions must be handled with care. Clinicians should be provided with clear guidelines on how to integrate model outputs into their decision-making process. The goal is to augment, not replace, clinical judgment with predictive analytics.

Finally, our work contributes to the broader field of healthcare analytics by demonstrating the feasibility and value of machine learning models in predicting critical medical conditions. The insights gained from this research pave the way for future studies to refine these models further and explore their integration into healthcare systems, ultimately aiming to improve patient outcomes through the power of data-driven decision-making.

5 CONCLUSIONS

Our endeavor to incorporate machine learning for the early detection of sepsis has yielded promising results, as evidenced by the performance of our K-Nearest Neighbors, Logistic Regression, and Random Forest models. Each model has demonstrated a unique strength in identifying sepsis from clinical datasets, emphasizing the potential of these methods to serve as valuable assets in clinical settings. The high degree of accuracy and balanced performance metrics suggest that these models can significantly contribute to the timely and accurate diagnosis of sepsis, potentially leading to improved patient outcomes and reduced healthcare burdens.

The KNN model's strong predictive accuracy suggests that it could be particularly useful as a screening tool in the early stages of patient care. The Logistic Regression model, while challenged by the class imbalance, provides understanding of the complicated nature of sepsis prediction. The Random Forest model stands out with its robust performance, highlighting the benefits of ensemble learning in handling complex patterns within high-dimensional data.

However, these models are not without their limitations. The challenge of class imbalance and the need for a reduction in false positives are areas that require further attentions. Future work should focus on refining these models through advanced algorithms, enriched feature engineering, and the adoption of ensemble techniques that can further enhance their predictive power. Moreover, the integration of machine learning predictions into the clinical workflow must be approached with careful planning, ensuring that these tools augment the expertise of healthcare professionals without superseding their critical role in patient care decisions.

In conclusion, this research has demonstrated the viability of machine learning approaches in the early detection of sepsis, offering a valuable step towards the integration of predictive analytics in healthcare. However, it also leaves room for future research to explore and expand upon. As we continue to refine these models and integrate them into clinical practice, we move closer to a future where the negative impacts of not just sepsis but a wide range of diseases can be substantially mitigated through the power of predictive analytics.

REFERENCES

[1] Gustavo EAPA Batista, Ronaldo C Prati, and Maria C Monard. 2005. Balancing strategies and class overlapping. In *International symposium on intelligent data analysis*. Springer, 24–35.

[2] Altman N. Krzywinski M. Bzdok, D. [n.d.]. Statistics versus machine learning. *Nat Methods* 15 ([n. d.]).

[3] Lynne Connelly. 2020. Logistic regression. *Medsurg Nursing* 29, 5 (2020), 353–354.

[4] Bishal Gyawali, Karan Ramakrishna, and Amit S Dhamoon. 2019. Sepsis: The evolution in definition, pathophysiology, and management. *SAGE open medicine* 7 (2019), 2050312119835043.

[5] Madiha Javeed, Ahmad Jalal, and Kibum Kim. 2021. Wearable sensors based exertion recognition using statistical features and random forest for physical healthcare monitoring. In *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*. IEEE, 512–517.

[6] Pavleen Kaur, Ravinder Kumar, and Munish Kumar. 2019. A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications* 78 (2019), 19905–19916.

[7] Suraj Kothawade, Pavan Kumar Reddy, Ganesh Ramakrishnan, and Rishabh Iyer. 2022. BASIL: Balanced Active Semi-supervised Learning for Class Imbalanced Datasets. *arXiv preprint arXiv:2203.05651* (2022).

[8] 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. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[9] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. 2020. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Critical care medicine* 48, 2 (2020), 210–217.

[10] 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, et al. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* 395, 10219 (2020), 200–211.

[11] James A Russell. 2006. Management of sepsis. *New England Journal of Medicine* 355, 16 (2006), 1699–1713.

[12] Mai Shouman, Tim Turner, and Rob Stocker. 2012. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology* 2, 3 (2012), 220–223.

[13] LAKSHYA SONI. 2022. *Sepsis prediction*. Kaggle. Retrieved December 5, 2023 from https://www.kaggle.com/code/lakshyasoni97/sepsis-prediction

[14] C. J. Van Rijsbergen. 1979. *Information Retrieval* (2 ed.). Butterworth-Heinemann.