Bridging Data and Clinical Insight: Explainable AI for ICU Mortality Risk Prediction

Ali H. Hassan[1], Riza bin Sulaiman[2], Mansoor Abdulhak[3], Hasan Kahtan[4]
Institute of IR 4.0 (IIR4.0), Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia[1, 2]
College of Computer and Cyber Sciences, University of Prince Mugrin, Madinah 41499, Saudi Arabia[1]
Computer Science Department at University of Oklahoma, Norman, Oklahoma 73019[3]
Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff CF5 2YB[4]

***Abstract*—Despite advancements in machine learning within healthcare, the majority of predictive models for ICU mortality lack interpretability, a crucial factor for clinical application. The complexity inherent in high-dimensional healthcare data and models poses a significant barrier to achieving accurate and transparent results, which are vital in fostering trust and enabling practical applications in clinical settings. This study focuses on developing an interpretable machine learning model for intensive care unit (ICU) mortality prediction using explainable AI (XAI) methods. The research aimed to develop a predictive model that could assess mortality risk utilizing the WiDS Datathon 2020 dataset, which includes clinical and physiological data from over 91,000 ICU admissions. The model's development involved extensive data preprocessing, including data cleaning and handling missing values, followed by training six different machine learning algorithms. The Random Forest model ranked as the most effective, with its highest accuracy and robustness to overfitting, making it ideal for clinical decision-making. The importance of this work lies in its potential to enhance patient care by providing healthcare professionals with an interpretable tool that can predict mortality risk, thus aiding in critical decision-making processes in high-acuity environments. The results of this study also emphasize the importance of applying explainable AI methods to ensure AI models are transparent and understandable to end-users, which is crucial in healthcare settings.**

***Keywords*—Explainable AI; healthcare; machine learning; predictive model**

I. Introduction

ICUs (Intensive Care Units) are specialized units that constantly monitor and treat patients with severe or potentially fatal illnesses. Due to the complexity of the ICU environments and the high-stake nature of patient care, accurate prediction of mortality risk in critically ill patients is a key aspect of adequate healthcare management [1], [2]. Machine learning algorithms and predictive models have shown great promise in addressing these needs by mining large amounts of clinical data to predict the risk of patient mortality. Predicting accurately has a significant effect on clinical decisions, allocation of resources, and the management of patients [3]. Machine learning methods have made tremendous progress and opened up new avenues for mortality prediction in the clinical world. To produce predictive insights, machine learning models are trained on complex datasets containing patient demographics, clinical measurements, and historical health records. Machine learning algorithms can explore more complicated relationships between variables than traditional statistical methods allow [4]. One example is through models utilizing extensive, large-scale electronic health record (EHR) data that has improved prediction performance for patient outcomes [5], [6].

Despite advances in machine learning, several challenges remain in accurately predicting mortality risk in ICU patients. For one, low-dimensional and high-dimensional complex healthcare data can increase the risk of overfitting and decrease the generalization performance of models if handled poorly [7], [8]. Also, the explainability of machine learning models is an important issue; medical professionals must have precise and reliable predictions to implement AI tools in clinical practice [9], [10], [11]. Black-box models can be opaque and difficult to interpret, preventing adoption in high-stakes medical contexts. Explainable Artificial Intelligence (XAI) frameworks, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), address this issue by providing transparent and interpretable explanations of model predictions, enhancing trust and usability [12], [13].

To mitigate these challenges for predictive models, the primary goal of this study is to develop an interpretable mortality risk prediction model incorporating machine learning algorithms with explainable artificial intelligence (XAI) methods. The dataset utilized in this study is derived from the 2020 Women in Data Science (WiDS) Datathon, which provides a rich profile of ICU patients, including various physiological and clinical variables [14]. The use of Explainable AI (XAI) methodologies, such as SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations), is intended to provide transparent insights into the model's decision-making process, thereby facilitating better understanding and trust in the predictive outcomes. This study addresses the limitations of previous studies by combining interpretability, accuracy, and clinical application, covering the way for reliable and actionable AI solutions in critical care settings.

The remainder of this study is structured as follows: Section II presents the background and motivation for the study. Section III provides a detailed description of the dataset's characteristics and explains the preprocessing methods employed in the study. Section IV presents the results of the training and evaluation of the AI model. Section V offers the discussion and considers the limitations of the study. Finally, Section VI concludes the paper,

summarizes the main points, and suggests areas for future research.

II. Background

In recent years, the healthcare industry has rapidly adopted machine learning (ML) and artificial intelligence (AI) approaches to construct predictive models that better decision-making and improve patient outcomes [15]. They excel at interpreting large amounts of medical data, including electronic health records (EHRs), genomic data, and medical images, revealing patterns that humans may overlook [16], [17]. Historically, interpretable methods like linear regression and naïve Bayes have been favored in healthcare domains such as neurology and cardiology, ensuring clarity for medical practitioners. Zhang et al. [18] developed an in silico prediction model for chemical-induced urinary tract toxicity using a naïve Bayes classifier, showcasing its effectiveness in toxicology assessments. Salman [19] explored heart attack mortality prediction by applying various machine learning methods, highlighting the potential of data-driven approaches to enhance clinical decision-making. Obeid et al. [20] introduced a deep learning model for the automated detection of altered mental status in emergency department clinical notes, offering insight into the capability of natural language processing in identifying critical health conditions.

However, while linear regression and naïve Bayes are interpretable by nature, these methods often struggle with complex or non-linear datasets. More recent approaches such as decision trees, K-nearest neighbors (K-NN), and support vector machines (SVMs) allow for more nuanced analyses of the data, which is particularly useful for output for conditions such as diabetes and heart diseases [21]. Abdalrada et al. [22] investigated machine learning models for predicting the co-occurrence of diabetes and cardiovascular disease, highlighting the potential for AI-driven techniques to discover overlapping risk factors and enhance early diagnosis. Karun [23] performed a comparative analysis of heart disease prediction algorithms, assessing the usefulness of several machine learning models, such as decision trees and SVMs, in increasing diagnostic accuracy. Rajkomar et al. [24] investigated the scalability and accuracy of deep learning using electronic health records, revealing how neural networks can analyze large volumes of patient data to improve prediction capacities while remaining adaptable across various medical situations.

While the field has advanced with deep learning algorithms that can deal with complex patterns in large datasets, their black-box nature presents a hurdle to clinical trust and uptake [12], [15], [25]. Interpretability is especially critical in supervised learning, commonly used in healthcare AI to predict specific health outcomes [26]. To counter the above problems, research has focused on developing XAI systems, which combine the complexities of the model and the ease of using it in a clinical setting [27], [28]. XAI controls the complexity of user explanations to those that a medical practitioner can understand, thus helping in forming opinions that can be acted on and trusted [29], [30], [31]. Furthermore, several evaluation methods for interpretability have been proposed, including application-grounded, human-grounded, and functionally grounded approaches. Application-based evaluations, which are performed while undertaking real-life activities along with the experts in the field, are particularly important for validating XAI models in the area of medicine [27]. Human-based evaluations involving laypeople are less expensive and easier to scale, but they may not capture the specific needs of medical practitioners [32]. Functionally grounded evaluations, which do not involve human users, assess interpretability based on proxy metrics like the complexity of decision trees, but they may lack practical relevance [33]. Even with these advances, there are still challenges like obtaining acceptable regulations, integrating them into existing frameworks, and teaching professionals in the field [25].

Despite these challenges, the demand for interpretable AI systems continues to grow as they offer the potential to revolutionize medical diagnostics and treatment planning by providing transparent, trustworthy, and actionable insights [32]. In this regard, this research aims to produce an explainable predictive model for predicting mortality risks in ICUs that use XAI and conventional AI techniques while overcoming inadequate past research in this field.

III. Method

A. Dataset Description

The dataset utilized in this study is sourced from the WiDS Datathon 2020 and comprises de-identified clinical data of 91,713 ICU patients collected over a year. This extensive dataset captures 186 physiological and clinical variables documented within the first 24 hours of ICU admission. The variables include a broad spectrum of demographic information, such as age, gender, and BMI, along with various health metrics like blood pressure, heart rate, and laboratory test results. These variables are essential for predicting patient outcomes, particularly mortality risk, in an ICU setting.

B. Data Preprocessing

1) Data cleaning: The first step in the data preprocessing involved loading the dataset and reducing the number of features from 186 to 28. This reduction was achieved by eliminating irrelevant or redundant features that did not contribute significantly to the prediction task.

2) Handling missing data: Columns containing significant missing data were eliminated to ensure the dataset's integrity and reliability. This step was crucial in ensuring the analysis was based on complete and accurate information. After removing these columns, the dataset was further refined, resulting in a final dataset consisting of 940 rows and 28 columns (Fig. 1, Fig. 2).

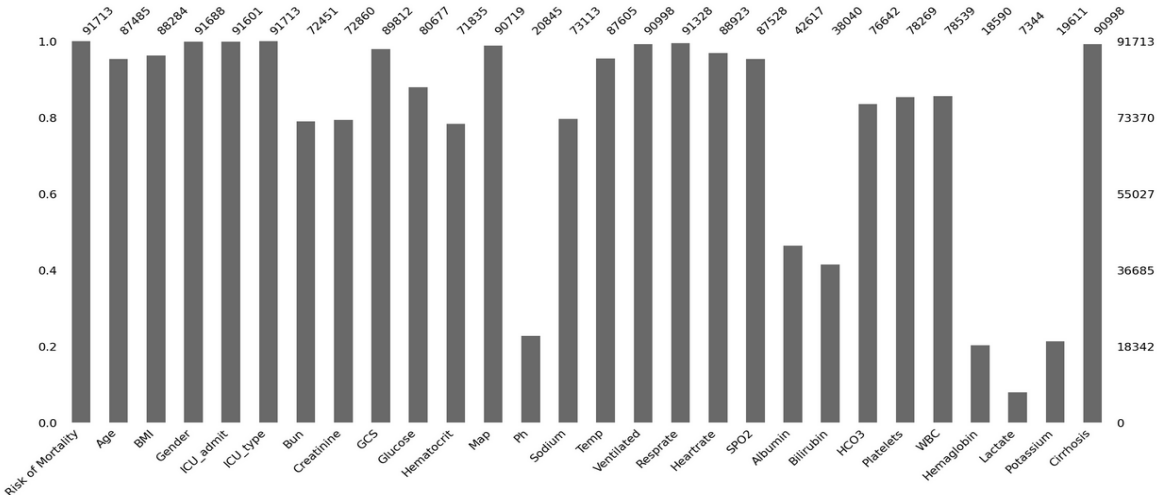

Fig. 1. Distribution of missing data across features.

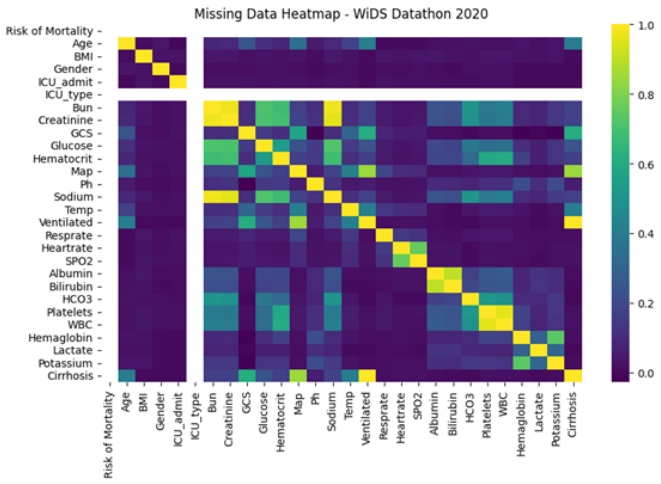

Fig. 2. Heatmap visualizing the correlations of missing values.

First, a correlation matrix of the missing values was generated, and a heatmap was used to identify patterns in the missing data. To address this, we applied the `dropna` function, which successfully removed the missing values and mitigated the impact of outliers. Rather than using imputation, we chose to eliminate entire columns with any missing data, prioritizing precision in healthcare predictions. While removing features with less than 5% missingness might seem counterintuitive, the critical nature of medical decisions made data completeness more important.

3) Data encoding: Data encoding techniques were applied to prepare the categorical variables for machine learning models. Specifically, Label Encoding was used to convert categorical variables such as Gender, ICU_admit, and ICU_type into numerical values, as illustrated in Fig. 3. This transformation allowed the models to process these variables effectively, as most machine learning algorithms require numerical input [34].

- Gender:
  - Original values: 'F', 'M'
  - Encoded values: 0, 1
- ICU_admit:
  - Original values: 'Floor', 'Operating Room / Recovery', 'Accident & Emergency', 'Other Hospital', 'Other ICU'
  - Encoded values: 0, 1, 2, 3, 4
- ICU_type:
  - Original values: 'Med-Surg ICU', 'MICU', 'CTICU', 'Neuro ICU', 'SICU', 'CCU-CTICU', 'Cardiac ICU', 'CSICU'
  - Encoded values: 0, 1, 2, 3, 4, 5, 6, 7

Fig. 3. Label encoding was applied to convert categorical variables.

4) Exploratory Data Analysis (EDA): EDA was conducted to uncover patterns and insights within the dataset. This analysis included visualizing demographic trends and identifying risk patterns among ICU patients. For instance, analysis of patients' age makeup clarified the age distribution. With concentrations in the 60–70 and 70–80 cohorts, age-related dynamics may impact mortality risk. Interestingly, the 50–60 cohort was equally large, demonstrating the sample's age diversity. Fig. 4 shows how the age distribution affects risk assessment. These insights guided the feature selection process and informed the subsequent steps in the analysis.

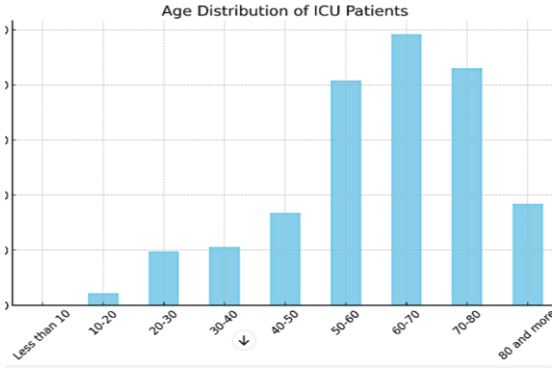

Fig. 4. Distribution of patient age groups.

5) Handling outliers: Outliers can distort the results of data analysis and model predictions. To address this, the Interquartile Range (IQR) technique was employed. Data points below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR were deemed outliers and deleted (Fig. 5, Fig. 6). This step ensured that the data used for model training was free from extreme values that could potentially bias the analysis [35].

6) Feature engineering: Feature engineering was performed to enhance the predictive power of the models. The SelectKBest method was used to identify the most significant features, and these selected features were then standardized using StandardScaler.

7) Balancing dataset: Given the class imbalance in the dataset, where high-risk cases were underrepresented, oversampling techniques were applied to balance the dataset. This step involved increasing the number of high-risk instances to ensure the model could accurately predict high-risk and low-risk cases. Balancing the dataset was vital for improving the model's sensitivity and reducing bias toward the majority class [36].

8) Data splitting: Partitioning the dataset into training and testing sets was the final step in data preprocessing. Eighty percent of the data was used to train the models, with the remaining 20% put aside for testing. This split meant that the models were evaluated using previously unseen data, resulting in a reliable assessment of their performance [37].

C. Model Training and Evaluation

The following six machine learning algorithms were chosen for evaluation: K-Nearest Neighbors, Random Forest, Decision Tree, Gradient Boosting, Logistic Regression, and Support Vector Machine. To optimize performance, GridSearchCV was implemented to tune hyperparameters for each model. The models were evaluated using F1-Score, Accuracy, Precision, and Recall classification measures. The best performing model was chosen for its proficiency in correctly forecasting ICU patient death.

D. Model Explainability

This study employed Explainable Artificial Intelligence (XAI) methods, particularly SHAP and LIME, to enhance model transparency. SHAP was used to interpret the importance of features in the Random Forest model, providing a detailed understanding of how each feature influenced risk predictions. LIME offered localized explanations for individual predictions, aiding in interpreting specific instances. After comparing both methods, SHAP was chosen as the primary tool for generating user-centric explanations, owing to its effectiveness in healthcare contexts [12], [38].

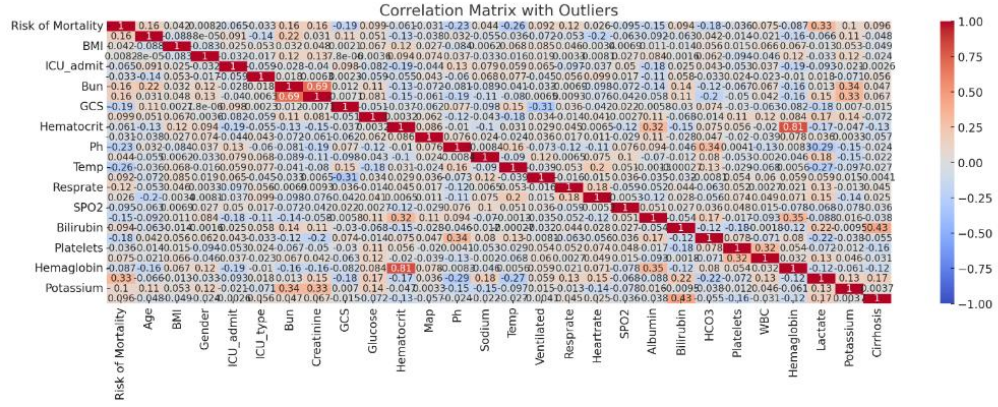

Fig. 5. Correlation matrix with outliers.

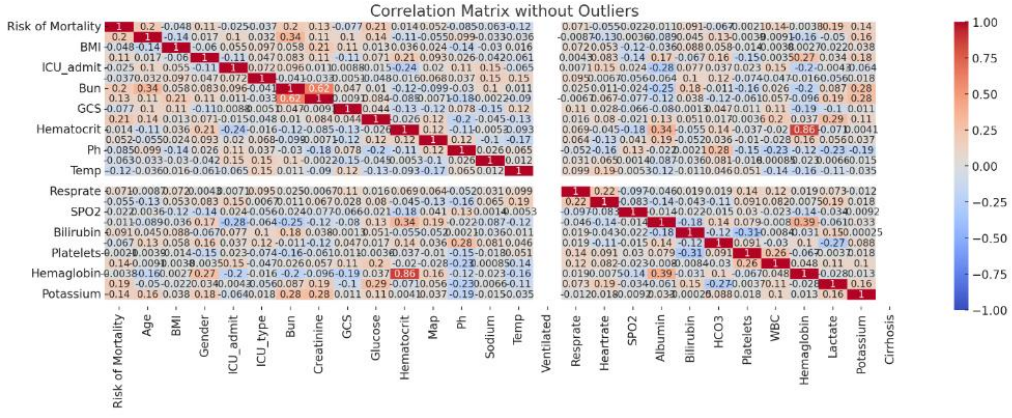

Fig. 6. Correlation matrix without outliers.

IV. Results

A. Model Performance

The performance of the six machine learning algorithms was compared based on their ability to predict mortality risk. As shown in Table I, the Random Forest algorithm outperformed the other models across all evaluation metrics, with an accuracy of 0.8511, precision of 0.8485, recall of 0.855, and an F1-Score of 0.8517.

TABLE I. Comparison of Machine Learning Techniques

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7252	0.7185	0.7405	0.7293
Random Forest	**0.8511**	**0.8485**	**0.855**	**0.8517**
Support Vector Machine	0.8321	0.8372	0.8244	0.8308
Gradient Boosting	0.8206	0.8088	0.8397	0.824
K-Nearest Neighbors	0.8244	0.7931	0.8779	0.8333
Decision Tree	0.7443	0.7133	0.8168	0.7616

B. Accuracy of Models

Fig. 7 illustrates the accuracy of the six machine learning algorithms. The Random Forest model proved the best accuracy, while the Support Vector Machine, K-Nearest Neighbors, and Gradient Boosting models performed comparably. In comparison, the Decision Tree and Logistic Regression models had slightly lower accuracy values.

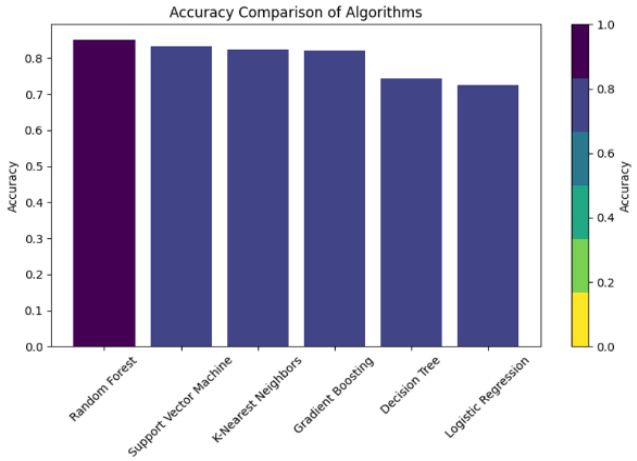

Fig. 7. Comparison of algorithms showing random forest model has the best accuracy.

1) ROC curve analysis: The ROC curve analysis emphasized the Random Forest model's superior performance. As shown in Fig. 8, the Random Forest model had the most significant AUC of 0.94, showing an excellent capacity to distinguish between high-risk and low-risk mortality cases. The Support Vector Machine model was next with an AUC of 0.92, while the Decision Tree model had the lowest AUC of 0.82.

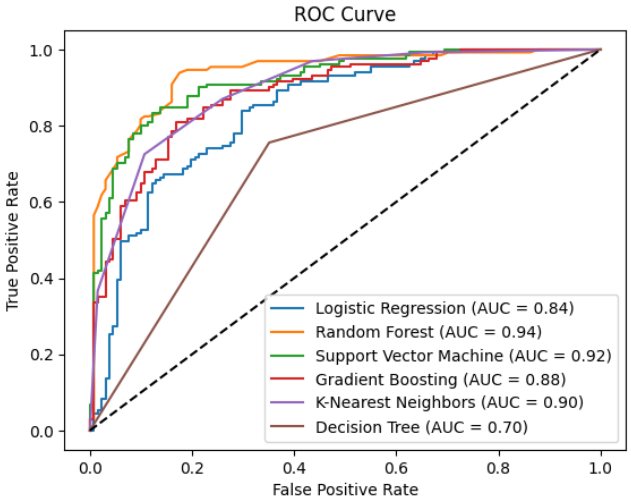

Fig. 8. ROC curve comparison for six machine learning algorithms.

2) Confusion matrix: The confusion matrix for each model provided insights into their classification performance. Fig. 9 illustrates the number of correct and incorrect predictions made by each model. The Random Forest model had the highest counts of true positives and true negatives, indicating its superior ability to classify both high-risk and low-risk mortality cases correctly.

C. Feature Importance Analysis

To enhance the interpretability of the model, Explainable Artificial Intelligence (XAI) methodologies, specifically SHAP and LIME, were employed. These techniques were valuable in making the model's decision-making process more transparent and understandable for healthcare practitioners.

1) SHapley Additive exPlanations (SHAP): SHAP was used to interpret the feature importance of the trained Random Forest model. A SHAP summary plot (Fig. 10) was generated using the TreeExplainer, which showed that features such as age, SOFA score, and lactate levels were significant contributors to the model's predictions. The SHAP values offered a clear grasp of how each characteristic affects mortality risk predictions, with greater SOFA scores and elevated lactate levels related to increased mortality risk, as acknowledged by studies such as [12], [38], [39].

2) Local Interpretable Model-agnostic Explanations (LIME): LIME was employed to explain individual predictions by generating local explanations for specific instances. Fig. 11 illustrates a detailed LIME explanation, showing the contribution of each feature to the prediction. For instance, one risk-predicting factor was high lactate level with significantly low temperature, with other factors like GCS, SPO2, ICU type, creatinine levels, age, and gender also playing roles. By breaking down predictions into these specific details, LIME enables healthcare practitioners to understand the rationale behind individual predictions, fostering trust in the model's outputs [13], [40], [41].

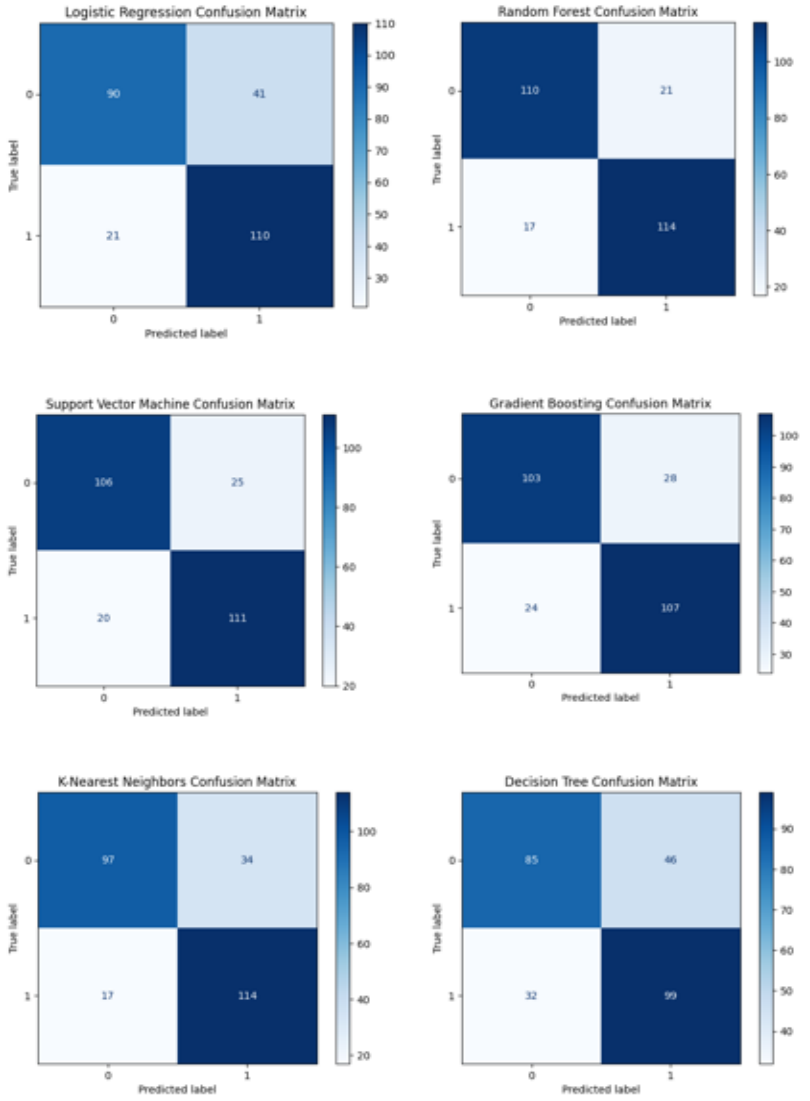

Fig. 9. Confusion matrix showing the number of correct and incorrect predictions made by each model.

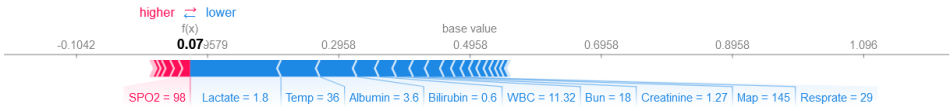

Fig. 10. SHAP summary plot.

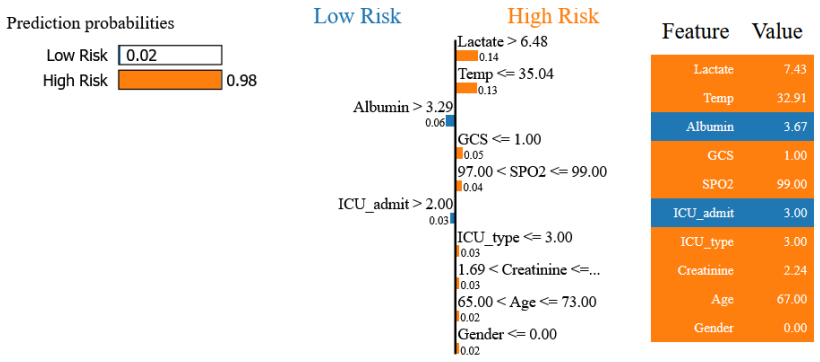

Fig. 11. LIME explanation.

V. Discussion

The results of this study are consistent with previous research that has demonstrated the effectiveness of Random Forest in predicting mortality risk in ICU patients. The Random Forest model's outstanding performance is primarily due to its capacity to handle large datasets and capture complicated correlations between variables. This study's findings align with previous research showing the usefulness of Random Forest in predicting mortality risk in ICU patients. The Random Forest model's exceptional success is primarily due to its ability to handle large datasets and identify the intricate connections between variables. The decision to balance the dataset by oversampling was crucial for improving the model's accuracy and reducing bias, a method supported by other studies, including that of [36]. Although [4] showed how effectively deep learning models capture non-linear correlations in complicated datasets, Random Forest was selected for this study because of its exceptional interpretability and widespread use in clinical settings. The use of SHAP and LIME provided valuable insights into the model's decision-making process. SHAP allowed for a global interpretation of feature importance, highlighting the significant role of clinical markers such as age, SOFA score, and lactate levels. These results are consistent with clinical expectations and established literature, thereby validating the model's predictions [12], [38], [39], [42]. The implications of this study for clinical practice are substantial. Implementing a Random Forest-based mortality risk prediction model enables healthcare professionals to more accurately identify patients at elevated risk of mortality and allocate resources more efficiently. This methodology may result in enhanced individualized care plans and superior patient outcomes. Furthermore, the model's ability to incorporate a wide range of clinical variables makes it adaptable to different ICU settings and patient populations.

While the results are promising, the study is not without limitations. The dataset is comprehensive; hence, the generalizability of the model may be restricted to different ICU populations. Despite the dataset is large, it only comes from one source. Hence, the generalizability of the model may be restricted to different ICU populations. This emphasizes the need for external validation over several datasets representing various geographical and clinical settings. Also, relying on these individual attributes raises concerns about potential biases. For example, demographic factors such as age or gender may potentially introduce systemic biases, limiting the model's generalizability across different patient populations [43]. Additionally, even though Random Forest performed more effectively than the other models in this study, investigating advanced methods—like ensemble approaches that combine deep learning with XAI tools—could improve the models' predictive abilities and interpretability.

VI. Conclusion

This study developed and evaluated a machine learning model to predict ICU patient mortality risk using the WiDS Datathon 2020 dataset. The Random Forest algorithm emerged as the most effective model, outperforming other algorithms in accuracy, precision, recall, and F1-Score. The AI mortality risk prediction model showed strong performance, with SHAP and LIME providing essential tools for enhancing model explainability. SHAP was particularly effective in offering clear and actionable explanations, making it the preferred method for developing user-centric explanations in this study. The study's findings highlight the potential of machine learning to improve clinical decision-making in ICU settings. Future research should focus on validating the model across multiple datasets from diverse geographic and clinical settings. Additionally, exploring integrating other machine learning techniques, such as deep learning, could further enhance the model's predictive accuracy.

References

[1] S. Barbieri, J. Kemp, O. Perez-concha, and S. Kotwal, “Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk,” Sci Rep, pp. 1–10, 2020, doi: 10.1038/s41598-020-58053-z.

[2] C. V Cosgriff, L. A. Celi, and D. J. Stone, “Critical Care, Critical Data,” Biomed Eng Comput Biol, vol. 10, p. 117959721985656, Jan. 2019, doi: 10.1177/1179597219856564.

[3] A. Hassan, R. bin Sulaiman, M. A. Abdulgabber, and H. Kahtan, “Balancing Technological Advances with User Needs: User-centered Principles for AI-Driven Smart City Healthcare Monitoring,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, p. 2023, 2023, doi: 10.14569/IJACSA.2023.0140341.

[4] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: Review, opportunities and challenges,” Brief Bioinform, vol. 19, no. 6, pp. 1236–1246, 2017, doi: 10.1093/bib/bbx044.

[5] S. M. Lauritsen et al., “Explainable artificial intelligence model to predict acute critical illness from electronic health records,” Nat Commun, vol. 11, no. 1, pp. 1–11, 2020, doi: 10.1038/s41467-020-17431-x.

[6] S. Boughorbel, F. Jarray, N. Venugopal, S. Moosa, H. Elhadi, and M. Makhlouf, “Federated Uncertainty-Aware Learning for Distributed Hospital EHR Data,” pp. 1–6, 2019, [Online]. Available: http://arxiv.org/abs/1910.12191

[7] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, “Developing clinical prediction models: a step-by-step guide,” BMJ, p. e078276, Sep. 2024, doi: 10.1136/bmj-2023-078276.

[8] M. A. Rahman Bhuiyan, M. R. Ullah, and A. K. Das, “iHealthcare: Predictive model analysis concerning big data applications for interactive healthcare systems,” Applied Sciences (Switzerland), vol. 9, no. 16, 2019, doi: 10.3390/app9163365.

[9] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible Models for HealthCare,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2015, pp. 1721–1730. doi: 10.1145/2783258.2788613.

[10] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nat Mach Intell, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[11] M. Cheng, X. Li, and J. Xu, “Promoting Healthcare Workers’ Adoption Intention of Artificial-Intelligence-Assisted Diagnosis and Treatment: The Chain Mediation of Social Influence and Human–Computer Trust,” Int J Environ Res Public Health, vol. 19, no. 20, 2022, doi: 10.3390/ijerph192013311.

[12] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Adv Neural Inf Process Syst, vol. 2017-Decem, no. Section 2, pp. 4766–4775, May 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, “Model-Agnostic Interpretability of Machine Learning,” no. Whi, Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.05386

[14] “WiDS Datathon 2020.” Accessed: Aug. 17, 2023. [Online]. Available: https://www.kaggle.com/c/widsdatathon2020/data

[15] M. A. Ahmad, A. Teredesai, and C. Eckert, “Interpretable machine learning in healthcare,” in Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, IEEE, Jun. 2018, p. 447. doi: 10.1109/ICHI.2018.00095.

[16] J. H. Chen and S. M. Asch, "Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations," New England Journal of Medicine, vol. 376, no. 26, pp. 2507–2509, Jun. 2017, doi: 10.1056/NEJMp1702071.

[17] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," 2019. doi: 10.1038/s41591-018-0300-7.

[18] H. Zhang, J. X. Ren, J. X. Ma, and L. Ding, "Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier," Mol Divers, vol. 23, no. 2, pp. 381–392, 2019, doi: 10.1007/s11030-018-9882-8.

[19] I. Salman, "Heart attack mortality prediction: An application of machine learning methods," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 27, no. 6, pp. 4378–4389, 2019, doi: 10.3906/ELK-1811-4.

[20] J. S. Obeid et al., "Automated detection of altered mental status in emergency department clinical notes: A deep learning approach," BMC Med Inform Decis Mak, vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12911-019-0894-9.

[21] D. D. Miller, "The Big Health Data–Intelligent Machine Paradox," Am J Med, vol. 131, no. 11, pp. 1272–1275, Nov. 2018, doi: 10.1016/j.amjmed.2018.05.038.

[22] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study," J Diabetes Metab Disord, vol. 21, no. 1, 2022, doi: 10.1007/s40200-021-00968-z.

[23] I. Karun, Comparative Analysis of Prediction Algorithms for Heart Diseases, vol. 1158. Springer Singapore, 2021. doi: 10.1007/978-981-15-4409-5_53.

[24] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," NPJ Digit Med, vol. 1, no. 1, 2018, doi: 10.1038/s41746-018-0029-1.

[25] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," Future Healthc J, vol. 6, no. 2, pp. 94–98, Jun. 2019, doi: 10.7861/futurehosp.6-2-94.

[26] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain Intelligence: Go beyond Artificial Intelligence," Mobile Networks and Applications, vol. 23, no. 2, pp. 368–375, 2018, doi: 10.1007/s11036-017-0932-8.

[27] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.08608

[28] R. Dwivedi et al., "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," ACM Comput Surv, vol. 55, no. 9, 2023, doi: 10.1145/3561048.

[29] A. Hassan, R. Sulaiman, M. A. Abdulgabber, and H. Kahtan, "Towards User-Centric Explanations For Explainable Models: A Review," Journal of Information System and Technology Management, vol. 6, no. 22, pp. 36–50, Sep. 2021, doi: 10.35631/JISTM.622004.

[30] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, "The relationship between trust in AI and trustworthy machine learning technologies," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 272–283, 2020, doi: 10.1145/3351095.3372834.

[31] L. Yu and Y. Li, "Artificial Intelligence Decision-Making Transparency and Employees' Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort," Behavioral Sciences, vol. 12, no. 5, 2022, doi: 10.3390/bs12050127.

[32] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," Electronics (Basel), vol. 8, no. 8, p. 832, Jul. 2019, doi: 10.3390/electronics8080832.

[33] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, Oct. 2018, pp. 80–89. doi: 10.1109/DSAA.2018.00018.

[34] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," J Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00305-w.

[35] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in Advances in Intelligent Systems and Computing, 2018. doi: 10.1007/978-981-10-7563-6_53.

[36] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," 2022. doi: 10.1016/j.compbiomed.2022.106043.

[37] Z. Ashfaq et al., "Embedded AI-Based Digi-Healthcare," Applied Sciences (Switzerland), vol. 12, no. 1, 2022, doi: 10.3390/app12010519.

[38] S. Das, M. Sultana, S. Bhattacharya, D. Sengupta, and D. De, "XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI," Journal of Supercomputing, 2023, doi: 10.1007/s11227-023-05356-3.

[39] N. Barr Kumarakulasinghe, T. Blomberg, J. Liu, A. Saraiva Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," in Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2020. doi: 10.1109/CBMS49503.2020.00009.

[40] R. K. Sheu and M. S. Pardeshi, "A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System," 2022. doi: 10.3390/s22208068.

[41] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions," J Comput Aided Mol Des, vol. 34, no. 10, 2020, doi: 10.1007/s10822-020-00314-0.

[42] S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, "Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare," Patterns, vol. 3, no. 1, 2022, doi: 10.1016/j.patter.2021.100395.

[43] R. R. Fletcher, A. Nakeshimana, and O. Olubeko, "Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health," Front Artif Intell, vol. 3, 2021, doi: 10.3389/frai.2020.561802.