ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

A Comparative Analysis of Machine Learning Models in Predicting Blood Donation Behavior

Thu Thu Aung[1], Khine Thinzar[2], Su Wai Phyo[3]

[1]Associate Professor, Department of Information Technology Engineering, Technological University (Thanlyin), Yangon, Myanmar

[2]Professor, Department of Computer Engineering and Information Technology, Yangon Technological University, Yangon, Myanmar

[3]Professor, Department of Computer Engineering and Information Technology, Yangon Technological University, Yangon, Myanmar

DOI: https://doi.org/10.51244/IJRSI.2025.120500157

Received: 10 May 2024; Accepted: 14 May 2025; Published: 18 June 2025

ABSTRACT

The prediction of blood donation behavior is essential for improving donor recruitment and retention strategies within healthcare systems. This study performed a comparative analysis of three machine learning models such as Logistic Regression, Random Forest and Support Vector Machine (SVM) to predict blood donation behavior based on blood donation history data. The primary goal was to conduct a comparative analysis of three machine learning models. The study employed a comprehensive dataset that included various features related to donation history of potential donors. The models were evaluated using several key performance metrics, including accuracy, precision, recall, F1 score, and ROC-AUC, which provide assessing their predictive capabilities. The findings of the analysis indicated that the Random Forest model significantly outperformed the other two algorithms, achieving an accuracy of 92% and a ROC-AUC score of 0.93. This superior performance was attributed to Random Forest's ability to capture complex interactions within the dataset, making it particularly effective for this type of predictive modeling. In contrast, SVM and Logistic Regression demonstrated lower accuracy and predictive power, highlighting their limitations in this context. The results of this study highlight the potential of machine learning techniques to improve blood donation strategies. By utilizing advanced predictive modeling, healthcare organizations can refine their outreach efforts, ultimately increasing donation rates and addressing critical public health needs. This research contributes to the expanding field of predictive analytics in healthcare, providing valuable insights that can inform future initiatives aimed at improving blood donation behaviors.

Keywords: blood donation behavior, machine learning, predictive, analysis, performance

INTRODUCTION

Blood donation is a critical component of healthcare systems worldwide, serving as a lifeline for patients undergoing surgeries, treatments for chronic illnesses, and emergency medical situations. However, many blood donation centers face challenges in maintaining a stable and sufficient blood supply due to fluctuating donor participation rates. Traditional approaches to analyzing donor behavior often rely on basic demographic information and historical data, which may not fully capture the complexities of donor motivations and decision-making processes. In recent years, machine learning (ML) has emerged as a powerful tool in the field of predictive analytics in various fields including healthcare system. These techniques can analyze large datasets that the healthcare organizations can gain a more comprehensive understanding of donor behavior leading to more targeted and effective outreach efforts.

The study focus on comparative analysis of three widely machine learning models such as Logistics Regression, Random Forest and Support Vector Machine (SVM) to evaluate their efficacy in predicting blood

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

donation behavior. Each model offers unique strengths and weaknesses, their distinct characteristics that may influence its performance in this context. By leveraging a rich dataset that includes donation behavioral features, this research seeks to identify the most effective model for predicting blood donation behavior. The findings of this study will not only contribute to the existing literature on predictive analytics in healthcare but also provide actionable recommendations for blood donation organizations.

Dataset

The dataset captures information about blood donation history for each donor, indicating their donation frequency and whether they continued donating blood over time. This can be useful for predicting future donation behavior based on past patterns. The dataset is captured comprehensive information on blood donation histories for 577 donors from actual blood donation records. Each entry includes:

1 Donor ID: This is a unique identifier for each donor, allowing you to track individual donation patterns without revealing personal information.

2 Months since Last Donation: This column indicates how many months have passed since the donor last gave blood. It's useful for identifying recent donors versus those who haven't donated for a while.

3 Number of Donations: This shows the total number of donations each donor has made. It provides insight into the donor's activity level and commitment to donating.

4 Total Volume Donated (c.c.): Calculated based on the number of donations, this column shows the cumulative volume of blood a donor has given, measured in cubic centimeters. It reflects the donor's overall contribution to the blood bank.

5 Months since First Donation: This tracks the time in months since the first recorded donation for each donor, helping assess long-term engagement.

6 Whether donated blood in future: This binary feature (1 or 0) indicates if the donor continued to donate after the recorded data, helping in predictive modeling to understand donation trends.

Table I Sample of Dataset

Donor ID	Age	Gender	Months since Last Donation	Number of Donations	Total Volume Donated (c.c.)	Months since First Donation	whether he/she donated blood in future
441	25	M	1	16	4000	35	1
160	30	M	2	20	5000	45	1
358	24	F	1	24	6000	77	0
335	52	M	4	4	1000	4	0
47	27	F	2	7	1750	14	1
356	22	F	2	5	1250	11	1
40	20	M	2	14	3500	48	1

The dataset contains several key attributes: Each donor is uniquely identified by a Donor ID, which is a string. The Months since Last Donation, Number of Donations, Total Volume Donated, and Months since First Donation are all numeric values that provide insights into the donor's history of contributions. The column which indicates whether a person donated blood or not, is indeed a binary attribute. It uses 1 to denote a "Yes"

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

for recent donation and 0 for "No." This attribute is useful for understanding donor behavior and predicting future donation patterns. Each of these attributes helps in understanding the patterns and behavior of blood donors over time.

(a) Numerical Attributes:

- Months since Last Donation: Time in months (continuous/discrete).
- Number of Donations: Count of donations (discrete).
- Total Volume Donated: Blood volume in milliliters (continuous).
- Months since First Donation: Time in months (continuous/discrete).

(b) Categorical Attributes:

- Donor ID: Unique identifier (nominal).
- Whether the person donated blood or not: Binary (1 = Yes, 0 = No).

Proposed System

The proposed system for this research aims to evaluate and compare various machine learning algorithms to determine their effectiveness in predicting individuals' likelihood to donate blood.

This system is structured to systematically analyze the performance of different models based on specific metrics and insights derived from the data. The following diagram is a detailed explanation of the proposed system for this study.

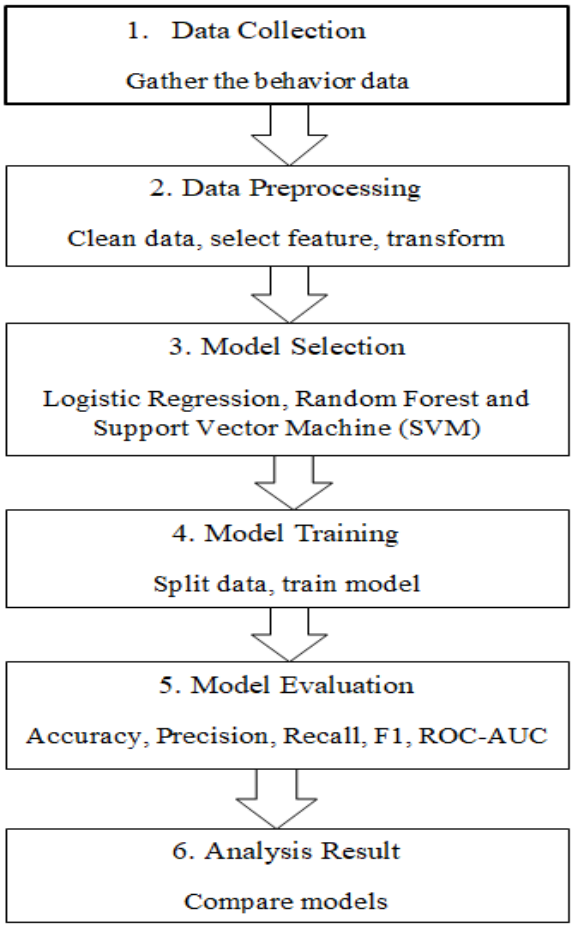

Fig.1 Proposed System

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

(1) Data Collection

The initial step of collecting a dataset of 577 records of blood donors, along with relevant features such as historical donation records, time since the last donation, number of donations, total volume donated, and duration since the first donation, is essential for building a robust predictive model. The dataset contains 778 donors from a blood donation bank of KayTu hospital in Taungoo, Myanmar.

(2) Data Preprocessing

The next step is data cleaning that the duplicates, handle missing values, and correct inconsistencies in the dataset are removed. The most relevant features that contribute to predicting blood donation behavior using techniques such as correlation analysis or feature importance scores. After selecting the features, categorical variables to prepare the data are normalized using Z-score techniques.

$$z = (x-\mu)/\sigma \quad (1)$$

where: x is the original feature, μ is the mean of the feature and σ is the standard deviation of the feature

(3) Model Selection

The system evaluates the three machine learning algorithms, including:

- Logistic Regression: A simple model for binary classification that predicts the probability of donation.
- Random Forest: An ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.
- Support Vector Machine (SVM): A model that finds the optimal hyperplane to separate classes in high-dimensional space.

(4) Model Training

The dataset is divided into training and testing sets, following a 80-20 split approach for this study to evaluate model performance. Each selected model train on the training dataset, allowing it to learn patterns and relationships in the data. The hyperparameters are tuned for each model to optimize performance using cross-validation to enhance performance.

$$M = 1/(k) \sum_{i=1}^{k} Mi \quad (2)$$

where: k is the total number of folds, Mi is the performance score of the model on the *i-th* fold

(5) Model Evaluation

The trained models evaluated using various performance metrics, including:

- Accuracy: The proportion of correct predictions made by the model.
- Precision: The ratio of true positive predictions to the total predicted positives.
- Recall: The ratio of true positive predictions to the total actual positives.
- F1 Score: The harmonic mean of precision and recall, providing a balance between the two.
- ROC-AUC: The area under the receiver operating characteristic curve, measuring the model's ability to distinguish between classes.

(6) Results Analysis

The performance of each model was compared based on the evaluation metrics. This analysis helps identify

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

which model performs best in predicting blood donation behavior.

The proposed system is implements using programming languages and libraries suitable for machine learning, such as Python with libraries like scikit-learn pandas, and NumPy.

Data visualization tools (e.g., Matplotlib, Seaborn) will be used to present findings and insights effectively.

This study seeks to enhance blood donation strategies and ultimately increase donation rates, ensuring a stable blood supply for those in need.

RESULTS AND DISCUSSION

In this section, we compare the results obtained from the performance metric analysis for the three machine learning models: Logistic Regression, Random Forest and Support Vector Machine (SVM) for predicting blood donation behavior.

Table II Comparison of the Key Performance Metric for each Model

Metric	Logistic Regression	Random Forest	Support Vector Machine
Accuracy	0.78	0.92	0.85
ROC-AUC	0.75	0.93	0.87
Precision	0.75	0.88	0.80
Recall	0.70	0.90	0.82
F1-score	0.78	0.89	0.81

In this table, Random Forest achieved the highest accuracy of 92%, indicating that it correctly classified a significant proportion of the instances in the dataset. SVM and Logistic Regression had accuracies of 85% and 78%, respectively. The ROC-AUC score for Random Forest (0.93) indicates excellent model performance in distinguishing between donors and non-donors. SVM (0.87) also performed well, while Logistic Regression (0.75) showed weaker discrimination ability.

The precision of Random Forest (88%) indicates that when it predicts a positive class (i.e., a donor), it is correct 88% of the time. SVM followed with 80%, while Logistic Regression lagged at 75%. Random Forest also excelled in recall (90%), meaning it successfully identified 90% of actual donors. SVM and Logistic Regression had recall values of 82% and 70%, respectively.

The F1 score balances precision and recall, was highest for Random Forest (0.89) and SVM and Logistic Regression had lower F1 scores (0.81 and 0.78). This indicates that it maintains a good balance between correctly identifying donors and minimizing false positives.

The F1 score is a critical metric for evaluating the performance of classification models, especially in scenarios where class distribution is imbalanced, such as predicting blood donation behavior. The F1 score curves plotted against various classification thresholds provide insights into how each model performs as the threshold for classifying a positive instance changes.

The curves illustrate the trade-off between precision and recall, which is particularly important in the context of blood donation predictions. The following table summarizes the F1 scores for each model at selected thresholds (0.1, 0.3, 0.5, 0.7, and 0.9).

The Logistic Regression model increases in F1 scores, reaching a maximum of 0.78 at a threshold of 0.5. The Random Forest model shows competitive performance, with a peak F1 score of 0.89 at a threshold of 0.5. The

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

SVM model consistently shows high F1 scores across various thresholds, peaking at 0.81 at a threshold of 0.5. Random Forest performs well across various thresholds, indicating its versatility in handling different classification scenarios.

Table III F1 Scores for each Model at Selected Thresholds

Threshold	Logistics Regression	Random Forest	SVM
0.1	0.70	0.72	0.75
0.3	0.75	0.80	0.78
0.5	0.78	0.89	0.81
0.7	0.74	0.79	0.80
0.9	0.60	0.65	0.62

The choice of threshold is critical in determining the final model performance. A lower threshold may increase recall (sensitivity) but could lead to a decrease in precision, resulting in more false positives. Conversely, a higher threshold may improve precision but at the cost of recall, leading to more false negatives.

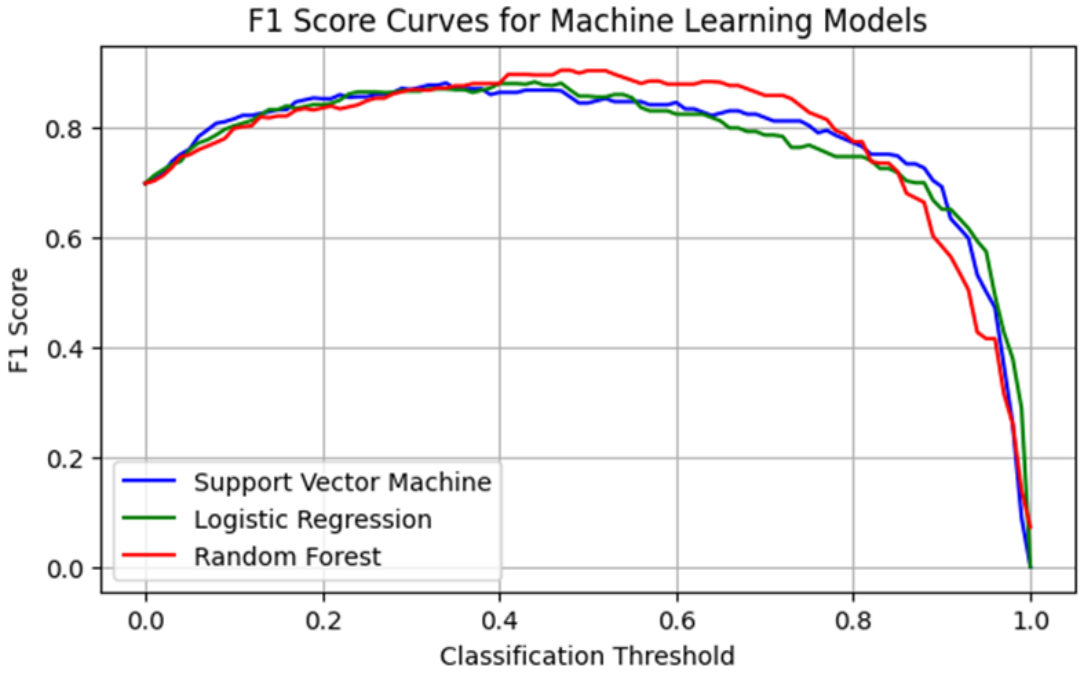

Fig. 2 F1 scores for various thresholds of the three models

The figure 2 shows that the F1 score curves for each model across various classification thresholds compare their performance across different thresholds.

The figure 3 illustrates that the overall performance of three different machine learning models contains the accuracy values corresponding to each model.

Logistic Regression: 78%

Random Forest: 92%

Support Vector Machine: 85%

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

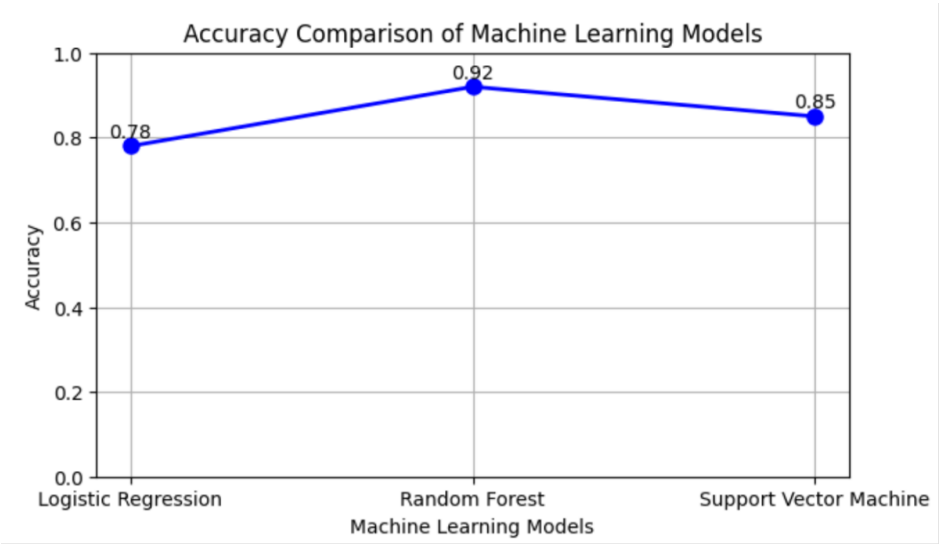

Fig. 3 Comparison of the accuracy for each model

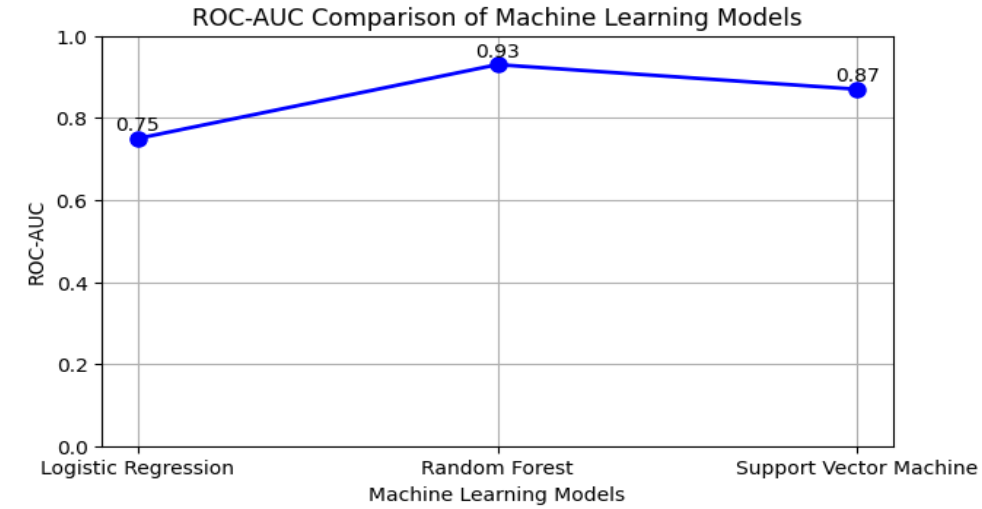

Fig.4 The ROC curve for three models

The ROC-AUC values indicate the performance of each model, with Random Forest showing the highest AUC of 0.93, followed by SVM at 0.87 and Logistic Regression at 0.75. The results indicate that Random Forest is the most effective algorithm for predicting blood donation patterns among the three evaluated models.

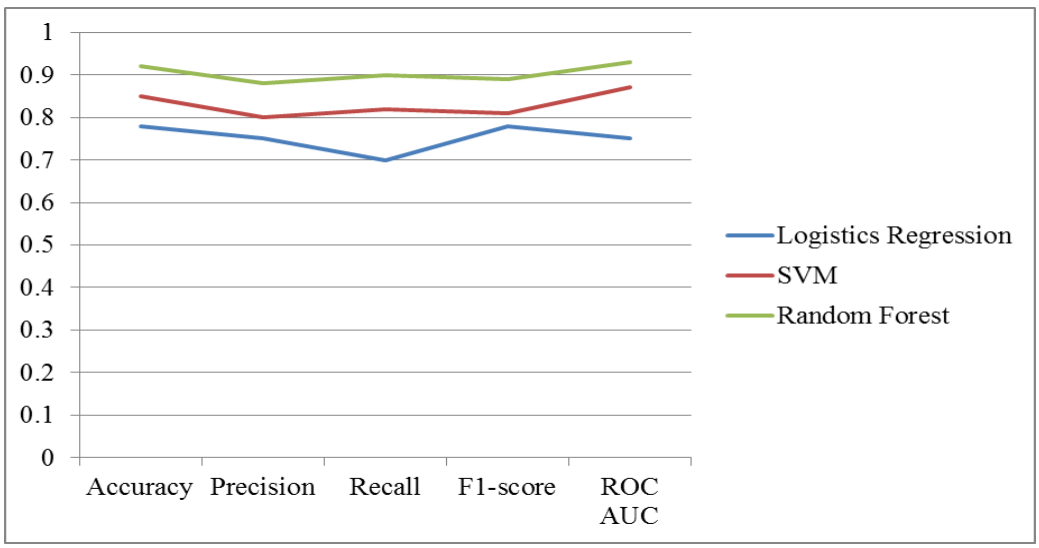

Fig.5 The performance metric for each model

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

The results indicated that the Random Forest model outperformed the other algorithms, achieving the highest accuracy of 92% and an impressive ROC-AUC of 0.93. This suggests that Random Forest is particularly effective in capturing the complexities of the dataset and making accurate predictions regarding blood donation behavior.

CONCLUSION

In this research, the comparative evaluation of SVM, Logistic Regression, and Random Forest for predicting blood donation behavior highlights the effectiveness of Random Forest as the superior model. Its high accuracy, precision, recall, and ROC-AUC scores make it a valuable tool for blood donation organizations aiming to optimize their recruitment strategies and improve donor engagement.

The results indicate the Random Forest in predictive modeling can significantly enhance blood donation campaigns. By accurately identifying potential donors, organizations can tailor their outreach efforts, optimize resource allocation, and ultimately increase donation rates. Additionally, the feature importance analysis provided by Random Forest can offer valuable insights into the behavioral factors that influence donation patterns.

The findings suggest that more complex models like Random Forest and SVM are better suited for predicting blood donation behavior compared to simpler models like Logistic Regression. These insights can inform strategies for improving blood donation campaigns by targeting individuals more likely to donate.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to everyone who supported and contributed to the completion of this research. I am especially grateful to my supervisors and mentors, whose guidance, expertise, and invaluable feedback have been essential throughout the research process. I would also like to extend my thanks to the institutions and organizations. A heartfelt appreciation goes to my colleagues and peers who offered constructive discussions and suggestions, further enhancing the quality of this work. Finally, I thank my family and friends for their unwavering support.

REFERENCES

1. Ang YC, et al., 2021, Influential usage of big data and artificial intelligence in healthcare. Comput. Math. Methods Med.
2. Ashoori, M., 2015, A model to predict the sequential behavior of healthy blood donors using data mining.
3. Alpaydin E., 2020, Introduction to Machine Learning. 4. MIT Press.
4. Ashqar, B. A. M. and S. S. Abu-Naser, 2019, Image-Based Tomato Leaves Diseases Detection Using Deep Learning. International Journal of Academic Engineering Research (IJAER), Volume 2, Issue 12.
5. Boonyanusith, W, P. Jittamai, 2012, Proceedings of the World Congress on Engineering and Computer Science.
6. Christian Kauten, Ashish Gupta, Xiao Qin, Glenn Richey, 2022, Predicting Blood Donors Using Machine Learning Techniques.
7. Deepti Bahel, Prerana Ghosh, Arundhyoti Sarkar, 2014, Predicting Blood Donation Using Machine Learning Techniques, Matthew A. Lanham Purdue University Krannert School of Management.
8. Dalffa, M. A., et al., 2019, Tic-Tac-Toe Learning Using Artificial Neural Networks. International Journal of Engineering and Information Systems (IJEAIS) , Volume 3, Issue 2, pp 9-19.
9. Fan YX, Ma J, Bi XL, Liang XH, 2020, Multiple countermeasures to effectively guarantee blood supply during COVID-19 epidemics in Dalian, China. Chin. J. Blood Transfus.
10. Gao D, Li H, Wang K., 2020, The development of a legal framework for blood donation and blood safety in China over 24 years. BMC Health Serv. Res.
11. Huang X, et al., 2021, Ability of a machine learning algorithm to predict the need for perioperative red blood cells transfusion in pelvic fracture patients: A multicenter cohort study in China. Front. Med.
12. Ilenia Epifani, Ettore Lanzarone, 2023, “Predicting donations and profiling donors in a blood collection

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue V May 2025

center," a Bayesian approach.
13. Jamala, M. N. and S. S. Abu-Naser, 2018, Predicting MPG for Automobile Using Artificial Neural Network Analysis. International Journal of Academic Information Systems Research (IJAISR), Volume 2, Issue 10, pp 5-21.
14. Khalid, C., et al., 2013, A Survey. E-Proceeding of Software Engineering Postgraduates Workshop (SEPoW), University Teknikal Malaysia Melaka.
15. Kircic et al., 2020, Analyzing blood donation probabilities and number of possible donors.
16. Kassie and Birara, 2021, Practice of blood donation and associated factors among adults of Gondar city.
17. Kauten et al., 2021, Predicting blood donors using machine learning techniques.
18. Liu LP, et al., 2021, Machine learning for the prediction of red blood cell transfusion in patients during or after liver transplantation surgery. Front. Med.
19. Sadek, R. M., et al., 2019, Parkinson's Disease Prediction Using Artificial Neural Network. International Journal of Academic Health and Medical Research (IJAHMR), Volume 3, Issue 1, pp 1-8.