AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND SOCIAL SCIENCES

Journal website: https://journals.must.ac.ke

A Publication of Meru University of Science and Technology

A systematic review of predictive blood donor retention models

Nahashon Kiarie[1*], Amos Chege Kirongo[1], Mary Mwadulo[1]

[1]*Meru University of Science and Technology, Meru, Kenya*

ARTICLE INFO

KEYWORDS

blood donor retention

machine learning

predictive model

healthcare

blood donation

ABSTRACT

Demand for blood and blood products is increasing due to population growth, medical advances, and increased disease. Availability of a stable blood supply is critical for healthcare organizations and requires effective donor recruitment and retention strategies. This systematic review paper examines the development and implementation of predictive models using machine learning techniques to classify and predict blood donor retention rates. The aim is to analyze the existing literature and provide insights into the design, performance and potential of such models through a systematic search of relevant databases. The reviewed studies include a variety of machine learning approaches and algorithms used to predict blood donor retention rates. These models use various demographic, behavioral, and historical donation data to predict the likelihood of a donor returning to donate blood. The utilization of machine learning techniques, such as decision trees, logistic regression, support vector machines, and neural networks, enables accurate predictions and enable healthcare organizations to implement targeted donor retention interventions to increase blood supply. The models' predictive performance reveals their capacity to recognize donors who are not likely to return and donate blood and target retention strategies appropriately, improving donor engagement and fostering long-term commitment. Several challenges and limitations face the identified existing models. They include the need for comprehensive and high-quality data, interpretability of complex models as well as the requirement for regular model updates to accommodate changing donor behaviors. There is need for development of versatile and comprehensive models with improved accuracy that can reduce the need for constant recruitment of new donors, which is costly and time-consuming enabling blood agencies to accurately predict donor retention rates, inform donor retention strategies, and prioritize resources appropriately and ultimately saving lives

Introduction

Background and Motivation

The demand for blood and blood products is constantly increasing due to population growth, advancements in medical procedures, and rising incidence of diseases such as cancer and chronic conditions that require regular transfusions. However, this increasing demand is not being met adequately, resulting in blood shortages and their subsequent impact on healthcare systems worldwide[1]. One of the primary reasons for blood donation scarcity is the insufficient recruitment and

*Corresponding author: Kiarie Nahashon Email: kiarienahashon12@gmail.com

https://doi.org/10.58506/ajstss.v2i2.206

AFRICAN JOURNAL OF SCIENCE, TECHNOLOGY AND SOCIAL SCIENCES ISSN :2958-0560

retention of blood donors. Blood donor retention is a critical factor in ensuring a stable and sustainable blood supply for meeting the healthcare needs of patients.

In many developing countries, up to 65% of blood transfusions are given to children under 5 years of age[2]. World health organization (WHO) recommends that the minimum country blood stocks at any time should be at least 1% of the population, but the average blood donation rate in Africa is 4.6 donations per 1000 population which is way below the recommended 10 donations per 1,000 people with some countries recording as low as 1.5 donations per 1000 people [3]. The age profile of blood donors shows that more young people donate blood in low- and middle-income countries than in high-income countries. In Kenya About 77% of blood donors are first-time donors and mostly from high schools and colleges. This provides large mass of young and healthy donors. It is therefore critical to put in strategies and measures to ensure that these donors are retained and are able to donate blood frequently throughout their lifetime.

This systematic review aims to explore the current landscape of blood donor retention models utilizing machine learning techniques. By examining the models used, variables considered, validation methods employed, and the strengths and weaknesses identified in the studies, this review aims to identify gaps and opportunities for further research.

Research Questions

The study was guided by these four research questions

i) What are the key machine learning algorithms used in predictive blood donor retention rate models?
ii) What are the most commonly used datasets for training and evaluating predictive blood donor retention rate models?
iii) How do different machine learning algorithms compare in terms of accuracy and performance in predicting blood donor retention rates??
iv) What are the main challenges or limitations of the existing predictive blood donor retention rate models using machine learning?

Methodology

This study conducted a systematic literature search to establish the existing research on predictive blood donor retention rate models. The process involved collecting, reviewing and in-depth analyzing relevant information on existing ML techniques for blood donor prediction and classification. Sources used in this study included Google Scholar, as well as scholarly databases such as Science Direct, IEEE and springer. Journal articles, books, and conference papers they were identified using a keyword-based search.

Application of machine learning in predicting blood donor retention

Globally there has been various applications of machine learning in blood donations. In their study on forecasting blood donor response using predictive modelling approach[4] used predictive modeling approach to predict whether a particular donor will donate blood within coming months. The research uses existing dataset obtained from the open database of Blood Transfusion Service Centre. The dataset contains five main variables that include: Recency which denotes the number of months since the person last donated blood, Frequency - total number of donations, Monetary - total blood donated in c.c.). Time in months since first donation and a binary variable representing whether the donor donated blood in March 2007. The study compares various classification algorithms such as decision trees, Artificial Neural networks and logistic regression. The results show that decision tree produced the best accuracy at 0.60. The accuracy achieved in this study was quite low and needs improvement additionally the variables used can be increased to improve on the accuracy.

The study by Zulfikar et al. (2018) classified eligibility of blood donors using decision trees and Naive Bayes classifiers. The study employed a data set of 500 Indonesian blood donors, of which 400 were used for training and 100 for testing[5]. The accuracy of the decision tree classifier was 78.5%, while the accuracy of the Naive Bayes classifier was 81.5%. The study utilized a real-world blood donor database and the research compared two distinct machine learning algorithms. The accuracy achieved by the algorithms needs improvement. The training dataset utilized for the study was also quite minimal which may limit generalization. Although Naïve bayes achieved a high accuracy in this study. The algorithm can struggle when dealing with imbalanced datasets and does not have inbuilt mechanism for handling missing values additionally the algorithm typically uses simple probabilistic models, making them less suitable for capturing complex relationships in data[6]. Decision trees utilized in the study are also susceptible to overlap, which can delay decision-making and increase memory consumption.

Using the dataset from Yangzhou Blood Station in China, Wu et al. 2022 gathered information about experienced blood donors recruited via short message service (SMS) and developed seven machine learning-based recruitment models[7]. Thirteen characteristics were outlined as a method for evaluating and predicting blood donors' intentions to donate. Area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F1 score were used to evaluate the performance of the prediction models on the complete dataset. Overall, 95,476 SMS recruitments and their donation outcomes were included in the modeling study. The accuracy of three superior models was validated using the tenfold cross-validation method based on blood groups. Blood donation interval, age, and donation frequency were discovered to be the most accurate predictors of the donation for experienced donors. The Extreme Gradient Boosting and Support vector machine models had the highest mean performance (95% CI) among the seven baseline models. Although the study utilized quite an extensive dataset and a variety of features, the study only dealt with only experienced blood donors leaving out students and new blood donors. Blood donors with missing critical data were also excluded from modelling processes, which affected the profile of the blood donors causing selection bias which may ultimately influence the results and limit generalization.

Pabreja et al. 2021 in their study on “Analytics Framework for Blood Donor Classification”, classified students from an Indian state university as potential blood donors or non-donors using data visualization techniques[8]. Students who were enrolled in a bachelor's degree program at Delhi state university, were surveyed using online questionnaire created with Google forms. Twenty questions regarding the characteristics of the donor were asked. Respondents were instructed to choose the option that best suited them on a Likert scale. Using convenient sampling technique, a total of 448 participants replied to the study. K-nearest neighbor and logistic regression algorithms were used to test the data on accuracy, precision, sensitivity and specificity. Recursive Feature Elimination (RFE) method was employed to produce a ranking of features and tenfold cross validation for validation. KNN classifier produced the best results with an Accuracy of 0.7027, Precision of 0.7209, Sensitivity value 0.7949, F1-score equal to 0.7561 and Specificity value 0.5789. The Results obtained in this study may be skewed since the data belongs to students only, the students also belong to the same university and same education level and hence related socio economic factors. The study utilized quite a minimal dataset which may limit generalizations moreover data provided via online questionnaires may not be completely verifiable, the accuracy achieved by the models also needs improvement.

Soft computing with data mining techniques have also found application for prediction in blood donation domain. Kewat et al 2018 employed Naive Bayes soft computing algorithm to classify and predict blood donors according to their sex and

blood group[9]. The blood donor's data was obtained from the Kota blood bank, comprising a total of 5656 cases and 12 attributes. The results showed that the generated classification rules carried out perfectly with accuracy rate 97.5588%. Despite the model's high accuracy in this study, Naive Bayes is a relatively simplistic probabilistic model, and this can be a disadvantage when dealing with complex datasets. The underlying premise of feature independence is one of the most major drawbacks of Naive Bayes. Given the class label, it assumes that all features are completely independent. This assumption is extremely simplistic and does not hold in many real-world settings. Naive Bayes is insensitive to feature relationships and dependencies because to its assumption of feature independence. When dealing with structured data, where feature interactions are important, this makes it less effective [6]. In addition, Naive Bayes is unable to handle missing data well and assigns non-zero probabilities to features that are not relevant, which can have an impact on the accuracy of predictions.

While predicting the return rate in young blood donors Cloutier, Et al (2021) extracted data from a blood donation management information system managed by Héma-Québec a non-profit organization that supplies hospitals in the Canadian province of Quebec with blood and other biological products of human origin[10]. The final dataset analyzed included 81 986 donors aged 18–24 at the time of their most recent donation. The data contained 11 main attributes, Additional information was acquired from the marketing database, that included data pertaining to donor contact details. The random forest model accurately predicted over 91% of donation frequencies, with an overall average error rate of 8.16% and specific error rates of 4.6% and 12.3% for the 'unreturned donor and returned donor groups respectively. The model's best predictive variables were found to be the number of marketing department contacts, the age of the donors, the number of adverse reactions during donation, the donors' status, and their ethnicity. Although the model achieved a considerable high level of accuracy, the study included only young donors between the age of 18- 24 years additionally donors that were contacted by the marketing professionals may result in bias. Random Forest is slow to construct and challenging to interpret, particularly when the ensemble comprises an extensive number of decision trees, as it must independently build and evaluate each decision tree[11]. Memory consumption can also be substantial when working with enormous datasets or ensembles containing numerous trees. This can restrict their applicability on systems with limited memory.

Selvaraj et al. 2022 while building a forecasting system for donation of blood using SVM Model, obtained data from a Blood Transfusion Service Center in Hsin-Chu City in Taiwan[12]. The dataset included 748 donors with five main variables: R (Recency - months since last donation), F (Frequency - total number of donations), M (Monetary - total blood donated in cc), T (Time - months since first donation), and a binary variable indicating whether a donor donated blood in March 2007 (1 for donating blood; 0 for not donating blood). The research utilized 10-fold cross validation. At 78.4 percent, Support Vector Classifier obtained the highest accuracy. The research was based on information from an isolated blood bank in India. It is possible that the study's findings may not be generalizable to other blood banks or countries. To aid the prediction model further, additional data tuples may be added. By adding more information, the results could become accurate. SVM can also be compared to other machine learning techniques to access the performance on the same dataset.

The study by Salazar-Concha et al. 2021 aimed to predict the intention to donate blood among blood donors using a Decision Tree Algorithm. [13]. The study utilized decision tree using C4.5, information gain for feature selection and tenfold cross validation. Face to face questionnaires were modelled based on theory of planned behavior and administered to adult users in two health centers in Valdivia (Chile). 197 participants respond-

ed, seven variables were used. The model achieved an accuracy of 84.17%. The sample size utilized in this study was quite minimal which can greatly limit generalization. The accuracy achieved in the study can also be can be improved

Various other studies have attempted to classify blood donors according to their characteristics and predict whether they are likely to donate in future. [14]Uses multiple logistic regression to identify the association between various blood donor characteristics such as the willingness of donate blood, number of months since the last donation, number of donations, total volume donated and the number of months since the first donation. Secondary data was retrieved from UCI Machine Learning Repository which gives information about blood donation by staff and students from university in Hsin-Chu City. The results showed that only three variables or factors are contributed to the willingness of donate blood which are: total months since last donation, frequency of donating blood and total volume of donated blood.

Discussion

In the systematic review of blood donor retention models using machine learning, several studies were examined to understand the current research landscape in this area. The reviewed studies employed various models, including logistic Regression, decision trees, random forests, support vector machines, and artificial neural networks, to predict blood donor retention. These models utilized a range of variables, including donor demographics, donation history, psychosocial factors, communication and engagement, external influences, health-related variables, and geographical factors. The validation methods employed included holdout validation, cross-validation, bootstrapping, and external validation.

Several gaps and opportunities for further research directions have been identified through the review. First, there is a need for standardized and consistent measurement of variables across studies. While the reviewed studies included various variables, the definitions and operationalization of these variables varied, making it challenging to compare and combine findings. Establishing a standardized set of variables and measurement techniques would enhance comparability and facilitate meta-analyses, enabling a more comprehensive understanding of blood donor retention [15]. Second, there is a need for more longitudinal studies that capture changes in donor behavior over time. Many of the reviewed studies relied on retrospective data, which limits the ability to understand the dynamics of donor retention. Longitudinal studies would provide insights into the temporal patterns of donor behavior, such as lapses and reactivations, and enable the identification of critical periods for intervention. Understanding the trajectory of donor retention over time would inform targeted strategies for improving retention rates[16].

Third, most reviewed studies focused on individual-level factors, neglecting the influence of social networks and community dynamics on donor retention. Exploring the role of social influence, social norms, and peer support in blood donor retention could uncover additional factors contributing to donor behavior. Additionally, investigating the impact of community-level variables, such as community engagement initiatives or local events, on donor retention would provide a more comprehensive understanding of the contextual factors influencing retention outcomes[17]. Fourth, while machine learning models were employed in the reviewed studies, there is a need for more advanced modeling techniques, such as ensemble methods and deep learning algorithms. Ensemble methods, such as stacking or boosting, have the potential to improve predictive accuracy by combining multiple models[18]. Deep learning algorithms, such as recurrent neural networks or transformers, could capture complex temporal dependencies in donor behavior and potentially enhance the prediction of donor retention. Exploring these advanced techniques could lead to more accurate and robust models for predicting and improving blood donor retention.

Fifth, there is an opportunity to leverage emerging data sources, such as electronic health records, social media data, or wearable devices, to enhance blood donor retention models. These data sources could provide additional insights into donor behavior and health-related factors that impact retention. Integrating diverse data types and employing advanced analytics techniques, such as data fusion or natural language processing, could uncover new predictors and improve the accuracy of retention models[19].

Finally, intervention studies are needed to evaluate the effectiveness of retention strategies informed by machine learning models. While the reviewed studies focused on prediction, the ultimate goal is to develop interventions that effectively improve donor retention rates[20]. Conducting randomized controlled trials or quasi-experimental studies to assess the impact of targeted interventions would provide valuable evidence on the practical applicability of machine learning models in real-world settings.

Conclusion

The potential of machine learning models in predicting blood donor retention has been demonstrated, albeit with differing levels of accuracy and trade-offs in terms of interpretability, computational complexity, and generalization capabilities. The datasets included a wide range of variables; however, there needed to be more clarity in data collection methods, the availability of comprehensive profiles, and the potential for biases in retrospective datasets. Future avenues for research should prioritize the enhancement of data quality and collection techniques, the resolution of missing data concerns, and the formulation of models that are both accurate and understandable.

References

[1] World Health Organization, "Global status report on blood safety and availability," 2017.

[2] C. M. Murtagh and C. Katulamu, "Motivations and deterrents toward blood donation in Kampala, Uganda," Soc Sci Med, vol. 272, p. 113681, Mar. 2021, doi: 10.1016/j.socscimed.2021.113681.

[3] "Blood Safety | WHO | Regional Office for Africa." Accessed: May 09, 2023. [Online]. Available: https://www.afro.who.int/health-topics/blood-safety

[4] C. Marade, A. Pradeep, D. Mohanty, and C. Patil, "Forecasting Blood Donor Response Using Predictive Modelling Approach," International Journal of Computer Science and Mobile Computing, vol. 8, no. 4, pp. 73–77, 2019.

[5] W. B. Zulfikar, Y. A. Gerhana, and A. F. Rahmania, "An Approach to Classify Eligibility Blood Donors Using Decision Tree and Naive Bayes Classifier," in 2018 6th International Conference on Cyber and IT Service Management (CITSM), IEEE, Aug. 2018, pp. 1–5. doi: 10.1109/CITSM.2018.8674353.

[6] N. Kalcheva, M. Todorova, and G. Marinova, "Naïve bayes classivier, decision tree and adaboost ensemble algorithm =- advantatges and disadvantages," 2020, pp. 153–157. doi: 10.31410/ERAZ.2020.153.

[7] H. Wu et al., "Predicting willingness to donate blood based on machine learning: two blood donor recruitments during COVID-19 outbreaks," Sci Rep, vol. 12, no. 1, p. 19165, Nov. 2022, doi: 10.1038/s41598-022-21215-2.

[8] K. Pabreja and A. Bhasin, "A Predictive Analytics Framework for Blood Donor Classification," International Journal of Big Data and Analytics in Healthcare, vol. 6, no. 2, pp. 1–14, Jul. 2021, doi: 10.4018/IJBDAH.20210701.oa1.

[9] A. Kewat and A. K. Sharma, "Evaluating the performance of naïve bayes classification algorithim FOR," J Emerg Technol Innov Res, vol. 5, no. 6, pp. 298–304, 2018.

[10] M. Cloutier, Y. Grégoire, K. Choucha, A. Amja, and A. Lewin, "Prediction of donation return rate in young donors using machine learning

models," ISBT Sci Ser, vol. 16, no. 1, pp. 119–126, Feb. 2021, doi: 10.1111/voxs.12618.

[11] D. A. Fife and J. D'Onofrio, "Common, uncommon, and novel applications of random forest in psychological research," Behav Res Methods, vol. 55, no. 5, pp. 2447–2466, Aug. 2022, doi: 10.3758/s13428-022-01901-9.

[12] P. Selvaraj, A. Sarin, and B. I. Seraphim, "Forecasting System for Donation of Blood Using SVM Model," Int J Res Appl Sci Eng Technol, vol. 10, no. 5, pp. 136–140, May 2022, doi: 10.22214/ijraset.2022.41940.

[13] C. Salazar-Concha and P. Ramírez-Correa, "Predicting the Intention to Donate Blood among Blood Donors Using a Decision Tree Algorithm," Symmetry (Basel), vol. 13, no. 8, p. 1460, Aug. 2021, doi: 10.3390/sym13081460.

[14] W. Hanieza, H. M. Sarkan, N. N. A. Sjarif, and Y. Yahya, "A Prediction Model for Blood Donation Using Multiple Logistic Regression," Open International Journal of Informatics (OIJI), vol. 7, no. 2, pp. 147–157, 2019.

[15] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," Journal of the American Medical Informatics Association, vol. 25, no. 10, pp. 1419–1428, Oct. 2018, doi: 10.1093/jamia/ocy068.

[16] L. Khomenko, L. Saher, and J. Polcyn, "Analysis Of The Marketing Activities In The Blood Service: Bibliometric Analysis," Health Economics and Management Review, vol. 1, no. 1, pp. 20–36, 2020, doi: 10.21272/hem.2020.1-02.

[17] A. Saad Alkahtani and M. Jilani, "Predicting Return Donor and Analyzing Blood Donation Time Series using Data Mining Techniques," 2019. [Online]. Available: www.ijacsa.thesai.org

[18] S. B. Golas et al., "A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data," BMC Med Inform Decis Mak, vol. 18, no. 1, p. 44, Dec. 2018, doi: 10.1186/s12911-018-0620-z.

[19] S. Campagnini, C. Arienti, M. Patrini, P. Liuzzi, A. Mannini, and M. C. Carrozza, "Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review," J Neuroeng Rehabil, vol. 19, no. 1, p. 54, Dec. 2022, doi: 10.1186/s12984-022-01032-4.

[20] C. Kauten, A. Gupta, X. Qin, and G. Richey, "Predicting Blood Donors Using Machine Learning Techniques," Information Systems Frontiers, vol. 24, no. 5, pp. 1547–1562, Oct. 2022, doi: 10.1007/s10796-021-10149-1.