ORIGINAL ARTICLE

Predicting Blood Donor Retention Using the Random Forest Classification Algorithm: A Machine Learning Approach

Marchel A. Acilador and Ann P. Opiña, PhD

Saint Louis University, Baguio City, Philippines

ABSTRACT

Background. Maintaining a stable and safe blood supply remains a persistent challenge for blood collection agencies, particularly in settings where donor participation is influenced by behavioral and psychological factors. Identifying reliable predictors of blood donor retention is essential for improving donor management and sustaining blood services.

Objective. To develop and evaluate a Random Forest–based machine learning model for predicting blood donor retention using demographic, behavioral, and psychological data.

Methods. A retrospective cohort with a cross-sectional exploratory component was conducted among blood donors from the Department of Health–Regional Blood Center of Cagayan Valley. Data were obtained from two sources: retrospective records of 612 donors (2023–2025), which included demographic characteristics, blood type, and donation history, and a survey administered to 100 prospective donors to assess psychological factors such as motivation, attitudes, satisfaction, and perceived barriers. Behavioral features—donation frequency, recency, tenure, and total donation count—were derived from donation records. Psychological variables were analyzed using Principal Component Analysis and K-means clustering. Model performance was evaluated using accuracy, F1 score, and area under the receiver operating characteristic curve (ROC–AUC).

Results. Demographic and biological characteristics, including age, sex, and blood type, described the donor population but were not key predictors of donor retention. Behavioral indicators—particularly donation frequency, recency, tenure, and cumulative donation count—were the strongest predictors of continued donation. Psychological analysis identified distinct donor profiles, with most respondents exhibiting high motivation and satisfaction, and smaller clusters reporting lower satisfaction or greater perceived barriers. The Random Forest model demonstrated strong predictive performance, with an accuracy of 0.989, F1 score of 0.994, and ROC–AUC of 0.994.

Conclusion. Blood donor retention is driven primarily by behavioral engagement and psychological factors rather than demographic or biological characteristics. The Random Forest model effectively identifies donors likely to return, supporting targeted donor engagement strategies and improved resource allocation in blood service facilities.

Keywords: donor retention, machine learning, Random Forest, behavioral predictors, psychological factors

Corresponding author: Marchel A. Acilador
Saint Louis University
Bonifacio St., Brgy. ABCR, 2600 Baguio City, Philippines
Email: 2235741@slu.edu.ph
ORCiD: https://orcid.org/0009-0004-9206-0680

INTRODUCTION

Blood donation remains a vital component of healthcare systems worldwide, ensuring the availability of blood products needed for emergency care, surgical interventions, and chronic transfusion therapies. However, maintaining an adequate and safe blood supply continues to pose significant challenges globally. The World Health Organization emphasizes the importance of recruiting and retaining regular, voluntary, unpaid donors, identified as the safest and most reliable contributors to national inventories.[1] Despite this, many countries—particularly low- and middle-income nations—struggle to meet recommended collection levels. In the Philippines, for example, efforts to achieve the WHO benchmark of obtaining blood from at least 1% of the population illustrate ongoing gaps in donor recruitment and, more critically, donor retention.[2] Studies further reveal that repeat donors contribute substantially to the total blood collected, with regular and repeated donors accounting for 81% of units in one national survey.[3] This heavy reliance on a shrinking pool of returning donors underscores the urgency of identifying factors that sustain donor participation, especially as recent reports indicate declining return rates among first-time donors.[4]

Understanding why donors fail to return has traditionally relied on psychological and behavioral models such as the Theory of Planned Behavior (TPB) and the Health Belief Model (HBM), which examine attitudes, subjective norms, motivations, self-efficacy, and perceived barriers.[5] Empirical findings highlight demographic factors, prior donation history, adverse reactions, altruistic motives, satisfaction with the donation experience, and convenience as influential predictors of donor behavior.[6-9] Other work suggests that intrinsic characteristics such as blood type may shape donation patterns, with type O individuals exhibiting higher donation likelihood due partly to awareness of their broader clinical compatibility.[10] Because demographic, behavioral, and biological attributes are routinely captured in donor information systems, they provide a rich foundation for understanding retention.[11] Nonetheless, traditional models are limited by their dependence on self-reported data and their inability to capture complex, nonlinear interactions among influencing factors.

As donor datasets expand, there is increasing urgency to adopt analytical approaches capable of integrating diverse variables and uncovering deeper patterns in donor behavior. Machine learning (ML) offers such potential, enabling the analysis of large, multidimensional datasets and revealing nonlinear structures that conventional statistical models may miss. Ensemble algorithms—including Random Forest (RF), Support Vector Machines, and gradient-boosting methods—have demonstrated strong predictive performance in modeling health behaviors.[7] Among these, Random Forest has gained prominence for its robustness, capacity to handle heterogeneous data, resistance to overfitting, and interpretability through feature importance measures. Studies have shown RF to outperform traditional approaches in predicting donation frequency, donor willingness, and return likelihood.[8,12,13]

In response to the pressing need for more effective donor retention strategies, this study aimed to develop and validate a Random Forest–based predictive model for blood donor retention among donors at the Department of Health–Regional Blood Center of Cagayan Valley over a three-year period (2023–2025). Specifically, the study sought to identify key demographic, behavioral, psychological, and biological predictors of donor return, and to evaluate the performance of the model using accuracy, F1 score, and area under the receiver operating characteristic curve (ROC–AUC) across training, validation, and test datasets. By integrating multiple dimensions of donor information, this study aimed to support data-driven strategies for improving donor retention and strengthening the stability of the blood supply.

METHODS

Research Design

This study employed a two-phase quantitative sequential design, utilizing a retrospective cohort with a cross-sectional exploratory component, to examine predictors of blood donor retention at the Department of Health–Regional Blood Center of Cagayan Valley. Retrospective donation records from 2023 to 2025 were analyzed to characterize donor demographics, behavioral attributes, and donation history, and to train a Random Forest model for predicting retention. All eligible donors with complete records were included in this phase. A cross-sectional survey was subsequently administered to a subset of donors during the 2025 collection period to assess psychological factors such as motivation, attitudes, satisfaction, and perceived barriers.

Study Population and Data Source

The study population consisted of registered blood donors at the Department of Health–Regional Blood Center of Cagayan Valley. From 2023 to2025, 6,301 donors were recorded, and those with initial and subsequent donations within this period were screened using predefined eligibility criteria for both the retrospective cohort and the prospective survey.

Retrospective Registry Cohort — Inclusion Criteria

Eligible participants were individuals who: (1) completed more than one whole blood donation between 2023 and 2025, allowing assessment of donor retention; (2) were 18–65 years old at their most recent donation; and (3) had complete, verifiable records. A total of 612 donors met these criteria.

Prospective Donor Survey — Inclusion Criteria

Eligible respondents were donors aged 18–65 years, medically cleared to donate, and with a recorded donation in 2023–2025, who provided informed consent. A total of 100 donors were invited to complete the survey.

Exclusion Criteria

For the retrospective cohort, donors were excluded for data inconsistencies or permanent medical deferral. For the prospective survey, donors unable to complete the questionnaire or those temporarily deferred (e.g., low hemoglobin, acute illness) were excluded.

Data Handling

The study utilized two distinct data sources: a retrospective donor registry comprising demographic and donation history (2023–2025), and a prospective survey evaluating psychological factors. To maintain strict compliance with the Philippine Data Privacy Act of 2012 and institutional anonymity protocols, survey responses were anonymized at the point of collection with no linkage established to individual retrospective IDs. Consequently, the retrospective dataset was used exclusively for descriptive analysis and machine learning model development, while the survey data underwent independent multivariate clustering. This approach provides a conceptual rather than a direct statistical bridge between the donor's psychological profile and their observed behavioral retention.

Missing data were addressed through data cleaning procedures prior to analysis. For the retrospective dataset, only donors with complete and verifiable records were included, and entries with missing or inconsistent values were excluded. For the prospective survey, questionnaires with incomplete responses were omitted from the analysis.

Operational Definition of Donor Retention

In accordance with DOH Administrative Order No. 2010-0002, a repeat donor is defined as an individual who has donated blood within one year prior to their current donation.[14] Consistent with this framework, donor retention in this study was assessed based on participation within a 12-month follow-up period after the initial recorded donation, as follows:

a. Retained Donor – A donor who returned to donate at least once within 12 months.
b. Lapsed Donor – A donor who did not return within the 12-month follow-up period.

Research Procedure

Ethical Approval

Ethical approval was obtained from the *Saint Louis University Research Ethics Committee (SLU-REC)* and the *Cagayan Valley Medical Center Research Ethics Committee (CVMC-REC)* prior to the initiation of any data collection, and the study was conducted in compliance with established ethical standards and the Data Privacy Act of 2012 (RA 10173).

Retrospective Data Collection (Phase 1)

Anonymized donation records from 2023 to2025 were obtained from the Department of Health–Regional Blood Center of Cagayan Valley. Extracted variables included donor demographics, blood type, donation dates, and donation frequency. All personally identifiable information was removed prior to extraction to ensure compliance with data protection standards. Behavioral features—recency, tenure, and frequency—were generated from historical donation patterns using Python-based feature engineering. The dataset then underwent standardized cleaning procedures, including variable normalization, removal of incomplete or inconsistent entries, and encoding of categorical variables through binary and one-hot encoding to prepare the dataset for machine learning analysis. The primary data attributes used in the study are summarized in Table 1. For descriptive analysis, age was categorized into groups to facilitate interpretation and comparison across donor subpopulations. However, for modeling purposes, age was retained in its original continuous form to preserve information and avoid loss of variability.

Prospective Survey Data Collection (Phase 2)

Prospective donors were recruited during mobile blood donation activities. Eligible individuals were informed about the study and provided written consent before completing a structured paper-based questionnaire. The survey assessed psychological determinants of donor retention, including

Table 1. Primary Data Attributes

Attribute Name	Attribute Type	Explanation	Predictor Category
Donor ID	Integer	Unique anonymized identifier assigned to each donor	Identifier
Sex	Categorical	Recorded as Male (M) or Female (F)	Demographic
Age	Continuous	Age of the donor at the time of the most recent donation	Demographic
Donation count	Discrete	Total number of donations recorded during the 2023–2025 period	Behavioral
First donation date	Date	Date of the donor's first recorded donation at the facility	Behavioral (History)
Donation dates	Date (List)	All donation dates recorded during the 2023–2025 period	Behavioral
Blood type	Categorical	Recorded ABO and Rh type (A+, A–, B+, B–, AB+, AB–, O+, O–)	Demographic

motivation, attitudes, satisfaction, and perceived barriers, using a five-point Likert scale. Completed questionnaires and electronic survey records were stored securely on encrypted devices and in locked physical storage, adhering to institutional and national data privacy regulations.

Model Development Dataset Preparation

The study employed the Random Forest classification algorithm to predict donor retention, selected for its ability to handle both categorical and continuous variables and model nonlinear relationships.

To prevent data leakage, an index date was defined. Behavioral predictors were computed using donation history prior to this date, while donor retention was assessed within the subsequent 12-month period, ensuring temporal separation between predictors and outcome.

Behavioral predictors were derived from historical donation records using Python (Pandas) based on donation dates and total donation count. Three key indicators were computed:

- ***Recency*** – number of days since the donor's last donation prior to the index date. It represents the number of days since a donor's last donation.
- ***Tenure*** – duration in days between the first and last donation prior to the index date, representing overall donor commitment.
- ***Frequency*** – average number of donations per month, calculated as total donations divided by tenure (in months)

These features captured donor engagement patterns and were included as key predictors in the Random Forest model. Their distributions were examined using histograms to assess variability.

Data Splitting

To prepare the dataset for modeling, the cleaned and feature-enhanced data were divided into three subsets: training, validation, and testing. A stratified 70/15/15 split was applied to maintain the proportional distribution of retained and lapsed donors across all subsets. The training set was used to develop the model, the validation set to tune hyperparameters, and the test set to evaluate performance on unseen data. The stratification process ensured that class imbalance did not bias model learning, maintaining fairness and generalizability in evaluation.

Model Training and Validation

The model training phase involved developing and optimizing a Random Forest Classifier to predict donor retention based on the retrospective dataset. The model was initially trained using default parameters to establish a baseline performance level. Accuracy, F1 score, and ROC-AUC were measured.

Hyperparameter Optimization

To confirm the model's stability and ensure that the baseline performance was not a result of overfitting, hyperparameter tuning was conducted using Grid Search Cross-Validation with three folds (cv=3) and F1-score as the optimization metric. The parameter grid explored values for the number of trees (n_estimators), maximum depth (max_depth), minimum samples per split (min_samples_split), minimum samples per leaf (min_samples_leaf), and number of features considered at each split (max_features).

Model Evaluation

The evaluation phase assessed the performance and generalization capability of the tuned Random Forest model using unseen test data, with model accuracy, F1-score, and ROC-AUC serving as the primary performance metrics. Classification outcomes were interpreted using standard definitions: true positives referred to retained donors correctly predicted as retained; false positives were donors who were not retained but were incorrectly classified as retained; true negatives were donors correctly identified as lapsed; and false negatives represented retained donors who were incorrectly predicted as lapsed.

Accuracy

Accuracy was generated by the Random Forest classifier using the built-in evaluation functions in Python. It represents the proportion of correctly classified cases—both retained and lapsed donors—relative to the total number of predictions made by the model.

F1 Score

The F1-score was computed automatically by the model using the harmonic mean of precision and recall. This metric provides a balanced assessment of the model's performance, particularly useful in the presence of class imbalance, by accounting for both false positives and false negatives.

ROC-AUC Score (Receiver Operating Characteristic – Area Under Curve)

The ROC-AUC score was calculated using Python's built-in functions to measure the model's ability to discriminate between retained and lapsed donors across all possible classification thresholds. An AUC value of 0.5 indicates no discriminative capability (equivalent to random guessing), while a value of 1.0 indicates perfect classification. Higher AUC values reflect a greater ability of the model to rank retained donors ahead of lapsed donors.

Performance Interpretation Criteria[15-17]

The following thresholds were used as general benchmarks for evaluating the acceptability of the model's performance. The evaluation criteria and corresponding acceptable thresholds for each performance metric are summarized in Table 2.

Table 2. Criteria Used for Evaluating Model Performance during the Training, Validation, and Testing Phases

Metric	Acceptable Threshold	Interpretation
Accuracy	≥75%	Indicates reasonably strong overall performance
F1 Score	≥70%	Reflects a good balance between precision and recall
ROC-AUC	≥0.80	Suggests strong discriminative ability

Exploratory Analysis of Psychological Factors Using PCA and K-Means Clustering

Psychological variables collected from the prospective survey were analyzed using exploratory, unsupervised methods. As the survey dataset originated from a separate sample and could not be ethically and operationally linked to retrospective donor records, these variables were not included in the Random Forest predictive model. Instead, Principal Component Analysis (PCA) and K-means clustering were applied to identify underlying psychological profiles that may inform donor retention behavior.

Data Preprocessing and Standardization

All survey items measuring motivation, attitudes, satisfaction, and perceived barriers were first inspected for completeness and consistency. Since Likert-scale items differ in variance, z-score standardization was applied to ensure equal contribution of each psychological variable to the clustering model. This prevented features with larger variance from dominating the analysis.

Principal Component Analysis (PCA)

Principal Component Analysis was used as a diagnostic tool to visualize the multidimensional structure of the psychological dataset. PCA reduced the high-dimensional feature space into two principal components, enabling visual assessment of potential grouping tendencies.

Clustering with K-Means

To identify latent donor profiles based on motivation, attitudes, satisfaction, and perceived barriers, a K-Means clustering algorithm was applied to the standardized psychological dataset. Multiple cluster solutions were evaluated using cluster stability, interpretability, and distribution of data points in PCA space.

The elbow method indicated that the optimal number of clusters lies between $k = 3$ and $k = 4$. Although PCA visualization suggested broader groupings, clustering was performed on the full multidimensional feature space, allowing for more refined distinctions among donors. Therefore, k = 4 was selected as it provided a better balance between model interpretability and the representation of nuanced psychological profiles.

Data Analysis

Data were analyzed using Python, integrating statistical methods and machine learning techniques. Descriptive statistics were used to summarize demographic and biological characteristics of donors.

Behavioral predictors of donor retention were examined using the Mann–Whitney U test to assess differences in donation count between retained and lapsed donors. Distributions of recency, tenure, and frequency were evaluated graphically, and relationships among behavioral features were assessed using correlation analysis.

The internal consistency of psychological constructs was evaluated using Cronbach's alpha. Psychological variables were analyzed separately using unsupervised analytical methods; Principal Component Analysis (PCA) was used for dimensionality reduction and visualization, followed by K-means clustering to identify distinct donor profiles. These analyses were exploratory and not included in the predictive model.

For predictive modeling, a Random Forest classifier was developed using demographic, behavioral, and biological variables. The dataset was split into training (70%), validation (15%), and test (15%) sets using a stratified approach. Model performance was evaluated using accuracy, F1 score, and ROC–AUC. Additional evaluation included confusion matrix analysis and feature importance ranking.

Ethics Approval and Consent to Participate

This study was conducted in accordance with established ethical standards for research involving human participants and data. Ethical approval was obtained prior to study commencement from the Saint Louis University Research Ethics Committee (SLU-REC) and the Cagayan Valley Medical Center Research Ethics Committee (CVMC-REC), in compliance with the Data Privacy Act of 2012 (RA 10173).

For the retrospective component, a Non-Disclosure Agreement was secured with the Department of Health–Regional Blood Center of Cagayan Valley. The requirement for informed consent was waived as all retrospective datasets were fully anonymized and de-identified prior to analysis, in accordance with institutional and national data governance guidelines.

For the prospective survey component, participation was voluntary, and written informed consent was obtained from all participants after they were informed of the study objectives, procedures, and data protection measures. Participants were assured of confidentiality and informed of their right to withdraw at any time without penalty.

RESULTS

Demographic, Behavioral, Psychological, and Biological Predictors

To identify demographic and biological predictors of blood donor retention within the study population, Table 3 is presented:

Table 3. Demographic and Biological Profile of Donors

Variables		Frequency (n=612)	Percent (%)
Sex	Male	330	53.9
	Female	282	46.1
Age Group (years)	18–24	93	15.2
	25–31	166	27.1
	32–38	125	20.4
	39–45	106	17.3
	46–52	73	11.9
	53–59	41	6.7
	60–61	8	1.3
Blood Type	A positive	158	25.8
	B positive	158	25.8
	AB positive	40	6.5
	O positive	256	41.8

The donor population (n = 612) was nearly sex-balanced (53.9% male, 46.1% female), with ages ranging from 18 to 61. The 25–31 age group was the largest (27.1%), followed by 32–38 (20.4%) and 39–45 (17.3%), while younger adults (18–24) represented 15.2%. Biologically, blood type O positive was most common (41.8%), followed by A and B positive (25.8% each).

To identify behavioral predictors of donor retention, the results are presented beginning with Table 4, followed by the visual analyses in Figures 1 to 3.

Table 4. Comparison of Donation Count between Retained and Lapsed Blood Donors Using the Mann–Whitney U Test

	Retained Median (IQR)	Lapsed Median (IQR)	U	Z	*p*-value
Donation Count	2 (2–3)	2 (2–2)	13,432.00	4.816	<0.001

Donation count differed significantly between retained and lapsed donors (p < 0.001), with retained donors showing greater variability (IQR: 2–3) than lapsed donors (IQR: 2–2).

As shown in Figure 2, recency displayed a right-skewed distribution, with many donors having donated recently and fewer with long intervals.

Tenure showed wide variability, indicating diverse donation histories.

Donation frequency was right-skewed, with a small subset of donors donating at high regularity

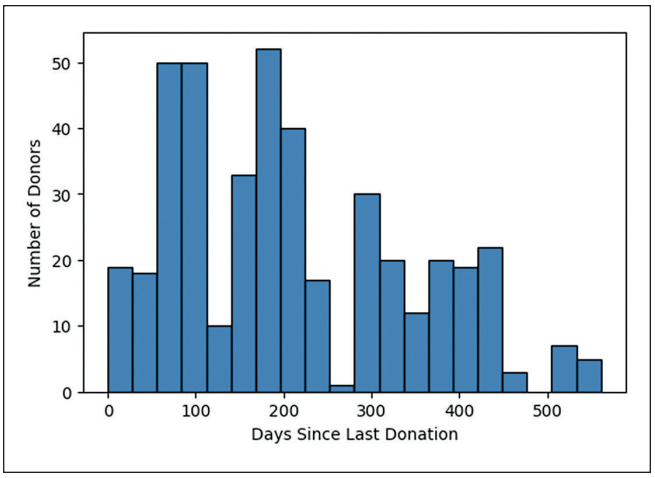

Figure 1. Distribution of donor recency.

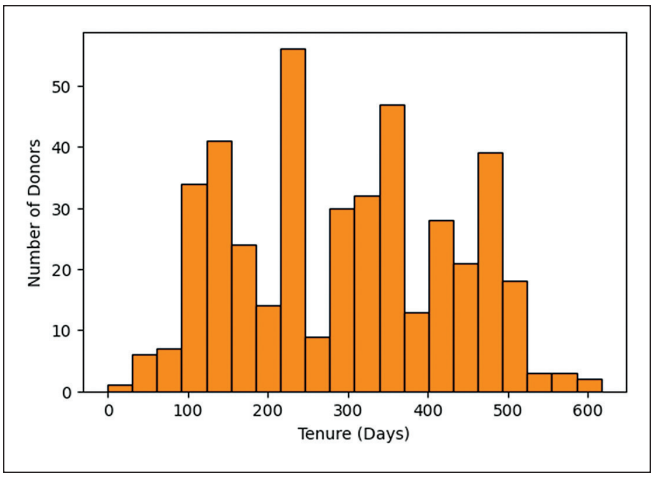

Figure 2. Distribution of tenure.

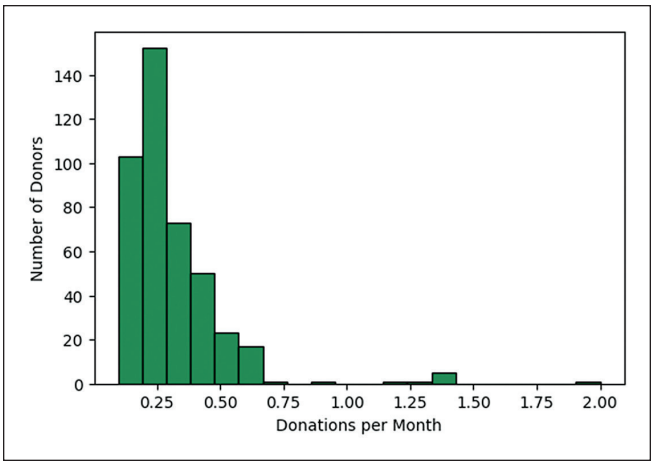

Figure 3. Distribution of frequency.

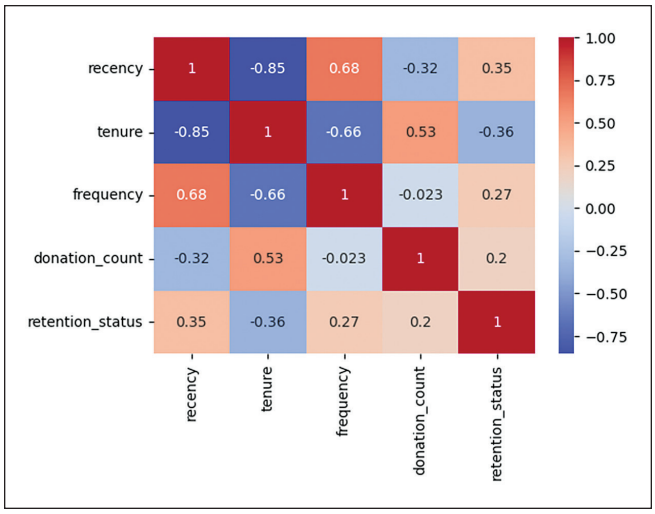

Figure 4. Correlation heatmap of behavioral features.

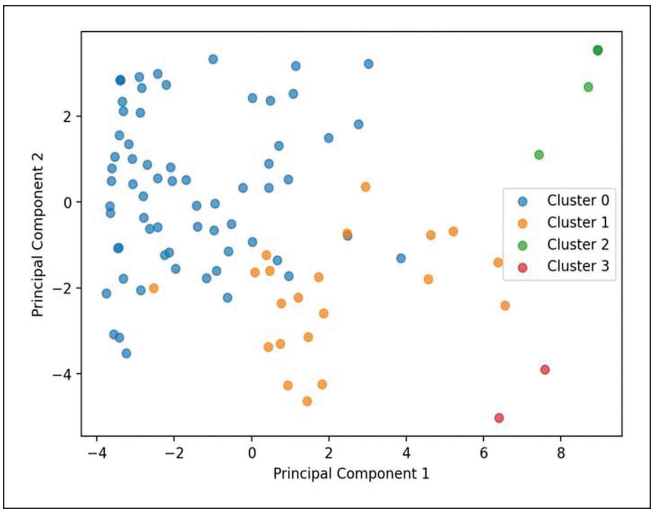

Figure 5. PCA visualization.

As part of the first objective, the study evaluated the predictive potential of key behavioral indicators by analyzing their correlations with one another and with retention status. Figure 4 displays the correlation heatmap used to determine feature relevance and guide preprocessing decisions for the machine learning model.

To establish the internal consistency of the psychological constructs to be used as potential predictors of blood donor retention, Table 5 is presented:

Table 5. Reliability Analysis (Cronbach's Alpha)

Construct	Cronbach's Alpha	No. of Items	Reliability Level
Motivation	0.749	8	Acceptable
Attitude	0.815	10	Good
Satisfaction	0.932	9	Excellent
Barriers	0.961	10	Excellent

All constructs demonstrated acceptable to excellent reliability (α = 0.749–0.961).

Model Training and Validation

Because the psychological responses from the prospective dataset could not be ethically or operationally integrated with the retrospective registry data for predictive modeling, it was expected that these variables would require independent analysis; therefore, dimensionality reduction using PCA and clustering via K-Means were employed to characterize underlying psychological patterns.

To examine the underlying psychological dimensions and identify distinct donor profiles that may serve as predictors of blood donor retention, and to inform feature preparation for the machine learning model, Figures 5 and 6 are presented:

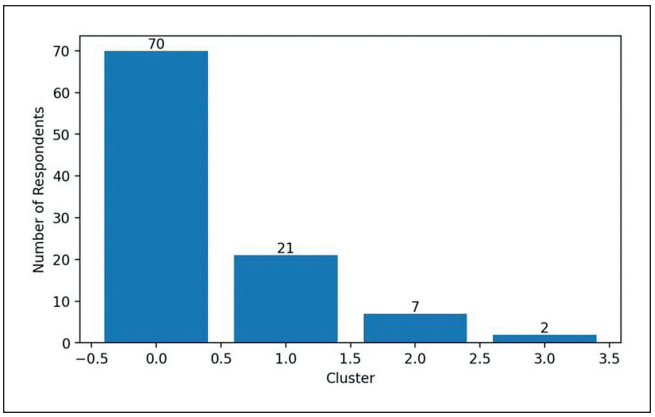

Figure 6. Predicted cluster distribution

This visualization confirmed that donors differed along latent psychological dimensions, supporting the applicability of clustering.

The final K-Means model produced four psychological donor clusters with varying group sizes: Cluster 0 comprised 70 respondents, Cluster 1 included 21 respondents, Cluster 2 contained seven respondents, and Cluster 3 consisted of two respondents.

To further interpret the psychological profiles represented by each cluster, Table 6 provides a descriptive characterization based on the observed patterns in motivation, attitude, satisfaction, and perceived barriers.

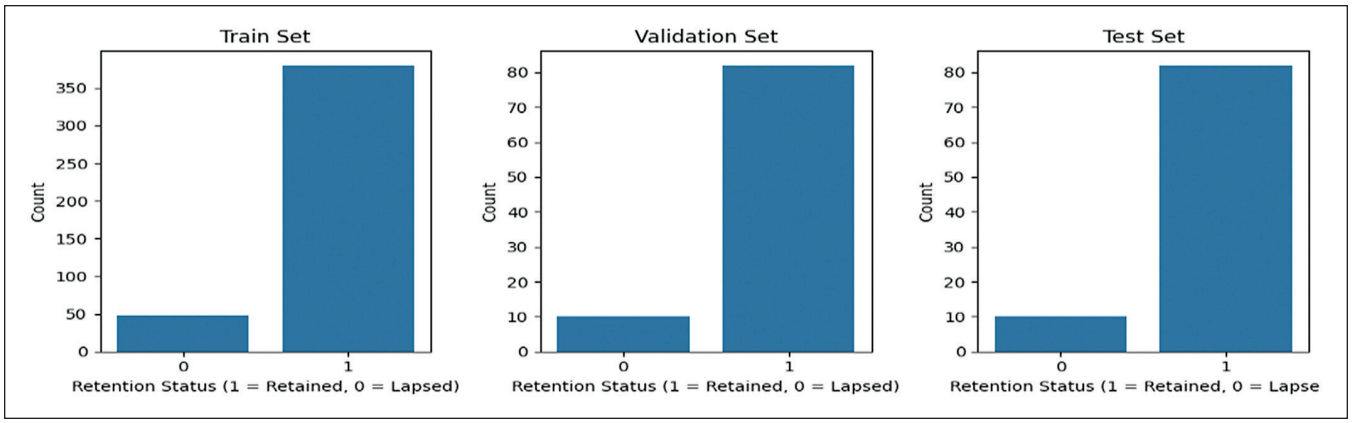

Figure 7. Class distribution across the training, validation, and test datasets.

Table 6. Interpretation of Psychological Clusters

Cluster	Meaning and Explanation
0	***Highly Motivated and Satisfied Donors*** Respondents who are characterized by strong positive attitudes, high satisfaction scores, and minimal perceived barriers.
1	***Moderately Engaged Donors*** Respondents who reflect neutral or slightly negative psychological scores
2	***Enthusiastic or Altruistic Donors*** Group with high motivational peaks and positive attitudes, indicating prosocial donors
3	***Low-Motivation, High-Barrier Donors*** Respondents who are characterized by low satisfaction, low motivation, and elevated perceived barriers. These respondents display psychological tendencies associated with lapsed or discouraged donors.

To present the distribution of donor retention status, defined as the outcome measure for identifying potential predictors of blood donor retention, Table 7 is shown.

Model Training and Performance Evaluation Using Accuracy, F1 Score, and ROC–AUC

The dataset was partitioned into training (70%), validation (15%), and test (15%) sets. Figure 7 shows the class distribution across all subsets.

The baseline and tuned Random Forest models demonstrated identical and consistently high predictive performance on both validation and test datasets, with near-perfect accuracy, F1-score, and ROC–AUC values (Table 8).

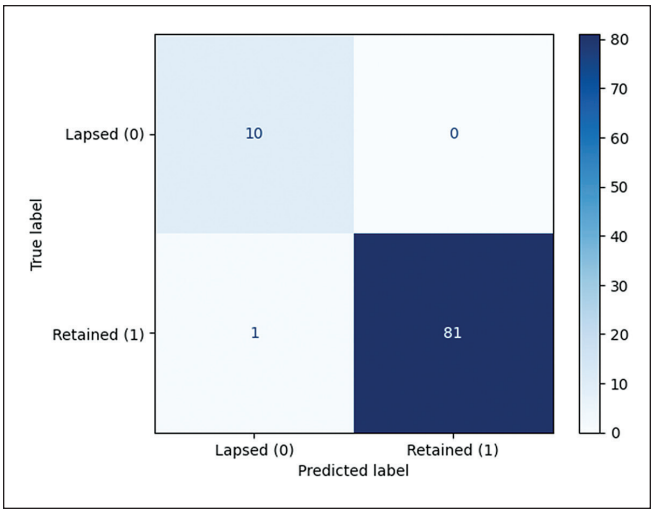

Figure 8. Confusion matrix of the final, trained model.

Figure 8 illustrates that the model correctly classified 91 out of 92 donors, with only one retained donor misclassified as lapsed.

Feature Importance Ranking

Figure 9 reveals that behavioral variables dominated the prediction process, with donation frequency emerging as the strongest indicator of donor retention.

Table 7. Frequency Distribution of Retention Status

Retention Status	Frequency	Percent
Lapsed	68	11.1
Retained	544	88.9
Total	612	100.0

A total of 88.9% of donors were retained.

Table 8. Performance of the Baseline and Tuned Random Forest Models

Model Phase	Dataset	Accuracy	F1-Score	ROC-AUC
Before Tuning (Baseline Model)	Validation	1.000	1.000	1.000
	Test	0.989	0.994	0.994
After Tuning (Optimized Model)	Validation	1.000	1.000	1.000
	Test	0.989	0.994	0.994

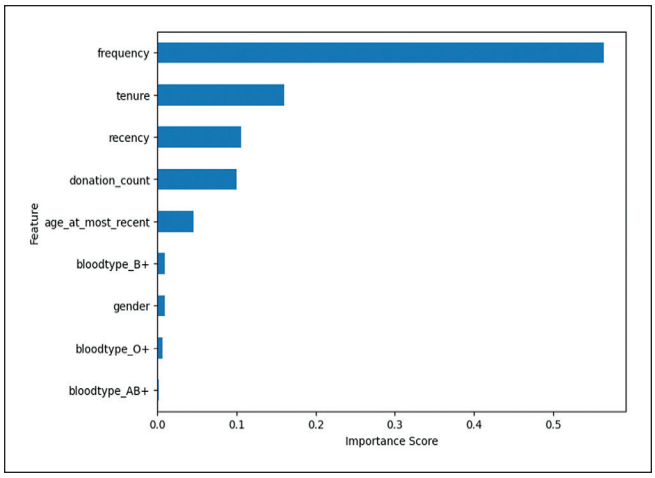

Figure 9. Feature importance ranking of the tuned model.

DISCUSSION

This study identified key predictors of blood donor retention using demographic, behavioral, psychological, and machine learning–based analyses, demonstrating that sustained participation is primarily influenced by behavioral and experiential factors rather than demographic or biological characteristics. The relatively balanced distribution across sex and age groups suggests equitable access to donation services within the study population. Although blood type O positive was most prevalent, consistent with national trends, biological attributes showed limited relevance in explaining retention behavior, supporting prior evidence that donor return is more strongly driven by behavioral and psychological determinants.[18,19]

Behavioral indicators emerged as the strongest predictors of donor retention. Donation count, frequency, recency, and tenure consistently reflected patterns of continued engagement, with higher donation frequency and greater cumulative donation count associated with increased likelihood of retention, while longer recency intervals were associated with decreased probability of return. These findings are consistent with behavioral theories of habit formation, wherein repeated participation reinforces donor identity and commitment.[20,21] From a public health perspective, these results indicate that maintaining regular donor engagement and minimizing gaps between donations are critical strategies for sustaining participation. Interventions such as automated reminders, structured follow-up systems, and regular scheduling of donation opportunities may help reduce lapsing and reinforce donor behavior.

Psychological constructs provided additional insight into the mechanisms underlying donor participation. While motivation and positive attitudes were generally aligned with retention-prone behaviors, donor satisfaction and perceived barriers appeared to play a more direct role in shaping donor experience. Higher satisfaction—particularly related to the quality of the donation process—was associated with continued participation, whereas increased perceived barriers, including logistical difficulties and negative procedural experiences, were linked to a higher likelihood of lapsing. These findings align with previous studies demonstrating that donor satisfaction is shaped by service quality and experiential factors.[19,21] However, it is important to note that psychological variables were not included in the predictive model due to the absence of a direct linkage between the survey and retrospective datasets. As such, their role in predicting donor retention could not be statistically established in this study and should be interpreted as exploratory.

Clustering analysis revealed distinct psychological donor profiles, ranging from highly motivated and satisfied individuals to smaller groups characterized by lower satisfaction and higher perceived barriers. Donors within favorable psychological profiles appeared more likely to align with retention-prone behaviors, whereas those experiencing barriers may be more vulnerable to discontinuation. Although some clusters were small, their identification highlights the importance of targeted interventions for at-risk donor subgroups, particularly those affected by negative donation experiences or logistical constraints. Tailored strategies addressing these barriers may yield meaningful improvements in retention outcomes.

The high donor retention rate observed in this study reflects the effectiveness of the region's community-based blood donation system. The integration of mobile blood donation drives within local government unit (LGU) programs, workplaces, and community activities reduces structural barriers such as distance and cost, while fostering a supportive and familiar environment for donors. Strong sociocultural norms of *bayanihan* and civic participation further reinforce sustained engagement. These findings are consistent with evidence showing that decentralized, community-integrated blood collection systems enhance accessibility and promote repeat donation.[21,22] Strengthening these systems remains a critical public health strategy for ensuring a stable blood supply.

Machine learning analysis further reinforced the dominance of behavioral predictors in donor retention. The Random Forest model demonstrated high predictive performance, with donation frequency, tenure, recency, and total donation count contributing most substantially to model predictions. The directionality of these predictors provides practical insight: frequent, recent, and sustained donation behavior increases retention likelihood, whereas declining engagement signals elevated risk of lapse. However, the high predictive performance should be interpreted with caution, as the dominance of behavioral variables and the class distribution may have contributed to model performance. Rather than serving solely as a classification tool, the model has practical value as a decision-support system, enabling early identification of donors at risk of discontinuation. This allows for the implementation of targeted, resource-efficient

interventions such as personalized reminders, follow-up engagement, and donor support programs.

This study has several limitations that should be considered in interpreting the findings. Selection bias may be present due to the non-random sampling approach used in the prospective survey, which may limit representativeness. Survivorship bias may also have occurred, as participants who responded to the survey are more likely to be active or engaged donors. There is a potential risk of overfitting given the relatively limited sample size and number of predictors; however, this was mitigated through stratified data splitting and cross-validation, which support model stability and generalizability. Data leakage was minimized by ensuring temporal separation between predictor variables and outcomes through the use of an index date. Additionally, due to data limitations, certain confounding variables were not included in the analysis, which may affect the robustness of the findings. The generalizability of the findings is also limited, as the study was conducted within a single regional blood center and may not fully represent donor populations in other geographic or healthcare settings.

Moreover, the findings demonstrate that blood donor retention is shaped by the interplay of behavioral continuity, donor experience, and accessibility of donation services. Effective public health strategies should therefore prioritize increasing donation frequency, reducing recency intervals, enhancing donor satisfaction, and minimizing perceived barriers. The integration of predictive analytics into donor management systems further enables proactive and targeted retention efforts, supporting the sustainability and resilience of blood supply systems. Future studies incorporating additional variables and larger, more diverse samples are recommended to further improve model generalizability and applicability.

CONCLUSION

This study identified the demographic, behavioral, psychological, and biological predictors of blood donor retention in the Cagayan Valley. It evaluated the performance of a machine learning–based model designed to predict which donors were most likely to return. Although demographic and biological characteristics—such as sex, age, and blood type—accurately described the donor population, they did not meaningfully influence retention outcomes. The balanced distribution of these attributes, together with the region's equitable LGU-led mobile blood donation system, aligns with prior research indicating that demographic factors contribute minimally to donor loyalty when access to donation opportunities is broad and community-driven.

Behavioral variables emerged as the most potent predictors of retention. Donation frequency, recency, tenure, and total donation count consistently distinguished retained donors from those who lapsed, reflecting stable participation patterns reinforced by regular mobile blood drives. These results support longstanding evidence that prior donation behavior is the strongest indicator of continued engagement.

Psychological factors also played a critical role in characterizing donor patterns. Moderate to strong correlations among motivation, attitudes, satisfaction, and perceived barriers revealed that donors differ meaningfully in their psychological orientations toward donation. The PCA and clustering analyses further confirmed the presence of distinct psychological donor profiles, with most donors exhibiting high motivation and satisfaction. At the same time, smaller groups reported lower satisfaction or perceived greater barriers. These findings highlight the importance of addressing donor experiences and perceptions to sustain long-term engagement.

The machine learning component demonstrated the operational value of integrating predictive analytics into donor management. The Random Forest classifier achieved exceptional stability and discriminative performance throughout validation and final testing. The near-identical pre- and post-tuning metrics, the high sensitivity and specificity observed in the confusion matrix, and the dominance of behavioral variables in feature importance rankings all point to the model's reliability and practical utility. By accurately identifying individuals likely or unlikely to return, the model provides a data-driven mechanism for optimizing outreach, minimizing unnecessary communication costs, and anticipating fluctuations in donor return rates. In settings where blood supply is vulnerable to seasonal and operational constraints, such predictive capability provides a meaningful advantage for planning and resource allocation.

If integrated into an existing donor management system, the predictive model could function as a decision-support tool to assist blood service personnel in retention planning. Using routinely collected donor data, the model could automatically generate retention risk classifications or probability scores at predefined intervals, enabling donors to be stratified according to their likelihood of return. These outputs could be presented through a simplified dashboard or alert system, allowing staff to identify donors who may benefit from targeted follow-up, scheduling assistance, or enhanced post-donation support. Rather than replacing current operational practices, the model would augment decision-making by providing timely, data-driven insights that support proactive and efficient donor engagement.

Data Availability Statement

The datasets generated and/or analyzed during the current study are not publicly available due to institutional data-sharing restrictions and the terms of the Non-Disclosure Agreement with the Department of Health–Regional Blood Center of Cagayan Valley.

Statement of Authorship

Both authors certified fulfillment of ICMJE authorship criteria.

Author Disclosure

Both authors declared no conflicts of interest.

Funding Source

None.

REFERENCES

1. World Health Organization. Blood safety and availability. World Health Organization. 2023.
2. Mappala ACA, Alican CAL, Dulay DCT, Mancita SCA, Utanes BYG, Clemente BM. Factors affecting voluntary blood donations among adults in Metro Manila, Philippines, as a basis for policy improvement on donor recruitment. Acta Med Philipp. 2023 May;57(5): 73-81. doi: 10.47895/amp.vi0.4351. PMID: 39678220; PMCID: PMC11635115.
3. Hashemi S, Maghsudlu M, Nasizadeh S, Esmaielifar G, Pourfathollah AA. Effective ways to retain first-time blood donors: a field-trial study. Transfusion. 2019 Sep;59(9):2893-2898. doi: 10.1111/trf.15392. PMID:31141337.
4. Association for the Advancement of Blood & Biotherapies. AABB24: State of the blood supply. 2024.
5. Bagot KL, Murray AL, Masser BM. How can we improve retention of the first-time donor? A systematic review of the current evidence. Transfus Med Rev. 2016 Apr;30(2):81-91. doi: 10.1016/j.tmrv.2016.02.002. PMID:26971186.
6. Salazar-Concha C, Ramírez-Correa P. Predicting the intention to donate blood among blood donors using a decision tree algorithm. Symmetry (Basel). 2021 Aug;13(8):1460. doi: 10.3390/sym13081460.
7. Wu HY, Li ZG, Sun XK, Bai WM, Wang AD, Ma YC, et al. Predicting willingness to donate blood based on machine learning: two blood donor recruitments during COVID-19 outbreaks. Sci Rep. 2022 Nov;12(1):19165. doi: 10.1038/s41598-022-24031-4. PMID:36348053; PMCID:PMC9644248.
8. Cloutier M, Grégoire Y, Choucha K, Amja AM, Lewin A. Prediction of donation return rate in young donors using machine-learning models. ISBT Sci Ser. 2021 Mar;16(1):119-126. doi: 10.1111/voxs.12626.
9. Kasraian L, Hosseini S, Dehbidi S, Ashkani-Esfahani S. Return rate in blood donors: a 7-year follow up. Transfus Med. 2020 Apr;30(2): 141-147. doi: 10.1111/tme.12676. PMID:31808280.
10. Sasaki S, Funasaki Y, Kurokawa H, Ohtake F. Blood type and blood donation behavior. SSRN Electron J. 2020 Jun 8. doi: 10.2139/ssrn.3171957.
11. Alkahtani AS, Jilani M. Predicting return donor and analyzing blood donation time series using data mining techniques. Int J Adv Comput Sci Appl. 2019 Aug;10(8):113-117. doi: 10.14569/IJACSA.2019.0100816.
12. Selvaraj P, Sarin A, Seraphim BI. Forecasting system for donation of blood using SVM model. Int J Res Appl Sci Eng Technol. 2022 May;10(5):136-140. doi: 10.22214/ijraset.2022.42068.
13. Kauten C, Gupta A, Qin X, Richey G. Predicting blood donors using machine learning techniques. Inf Syst Front. 2022 Oct;24(5): 1547-1562. doi: 10.1007/s10796-021-10137-2.
14. Department of Health. Administrative Order No. 2010-0002: Policies and guidelines pertinent to the establishment and operation of local blood councils to support the implementation of the National Voluntary Blood Services Program for blood safety and adequacy, quality care and patient safety. Manila (PH): Department of Health; 2010.
15. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009 Jul;45(4): 427-437. doi: 10.1016/j.ipm.2009.03.002.
16. Powers DMW. Evaluation: Precision, recall, F-measure, ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011 Jan;2(1):37-63.
17. Kuhn M, Johnson K. Applied predictive modeling. New York (NY): Springer; 2013. doi: 10.1007/978-1-4614-6849-3.
18. Kowalsky JM, France CR, France JL, Whitehouse EA, Himawan LK. Blood donation fears inventory: development and validation of a measure of fear specific to the blood donation setting. Transfus Apher Sci. 2014 Oct;51(2):146-151. doi: 10.1016/j.transci.2014.08.007. PMID:25218310.
19. Bednall TC, Bove LL. Donating blood: a meta-analytic review. Transfus Med Rev. 2011 Oct;25(4):317-334. doi: 10.1016/j.tmrv.2011.03.002. PMID:21641781.
20. Lemmens KPH, Abraham C, Hoekstra T, Ruiter RAC, De Kort WLAM, Brug J, et al. Why don't young people volunteer to give blood? An investigation of the correlates of donation intentions among young nondonors. Transfusion. 2005 Jun;45(6):945-955. doi: 10.1111/j.1537-2995.2005.04379.x. PMID:15859912.
21. Masser BM, White KM, Hyde MK, Terry DJ, Robinson NG. Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model. Transfusion. 2009 Feb;49(2):320-329. doi: 10.1111/j.1537-2995.2008.01981.x. PMID:18954389.
22. Ferguson E, Atsma F, de Kort W, Veldhuizen I. Exploring the pattern of blood donor beliefs in first-time, novice, and experienced donors: differentiating reluctant altruism, pure altruism, impure altruism, and warm glow. Transfusion. 2012 Feb;52(2):343-355. doi: 10.1111/j.1537-2995.2011.03279.x. PMID:21790699.