



International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Comparative Study of Machine Learning Algorithms for Blood Donor Availability

K. Manicka Dharsini, Dr. Nancy Jasmine Golden

Department of Computer Applications, Sarah Tucker College, Tirunelveli-7.

ABSTRACT

Efficient identification of available blood donors is an important requirement in emergency healthcare situations, where delays in locating suitable donors can affect timely blood availability. This paper presents a comparative analysis of four supervised machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN), for predicting blood donor availability. Donor demographic, health, location, and donation history attributes are considered as predictive features. The dataset is preprocessed by selecting relevant attributes and converting categorical values into numerical representations before model training. The four algorithms are trained and evaluated under the same experimental conditions using Accuracy, Precision, Recall, and F1-Score. The comparative results indicate that Random Forest achieves the best overall performance among the evaluated algorithms. The trained machine learning approach is integrated with location and blood group-based donor search to identify and prioritize potentially available donors matching the required criteria. The proposed approach provides a data-driven mechanism for improving donor identification and supporting efficient blood donor selection during emergency healthcare situations.

Keywords: Blood Donor Availability, Machine Learning, Donor Prediction, Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Comparative Analysis, Classification.

1. INTRODUCTION

Blood donation plays an important role in emergency healthcare, where timely identification of suitable donors is necessary for patients requiring blood during accidents, surgeries, severe blood loss, anemia, and other medical conditions. Registered donor records alone do not ensure availability, as it can change due to previous donation activity, medical conditions, personal circumstances, and the time since the last donation. Therefore, donor identification is not only a record retrieval problem but also a prediction problem involving multiple donor characteristics.

Traditional donor search methods mainly use fixed attributes such as blood group and location. Although they identify matching donors, they do not determine whether a donor is likely to be available when required. Manual verification may be needed when many records are obtained, while outdated information can reduce search reliability. These limitations indicate the need for an approach that analyses donor information beyond simple filtering.

Machine learning provides an alternative by learning patterns from recorded donor data. The proposed paper uses demographic, health, location, and donation-history information to identify patterns associated with donor availability. Analysing these attributes together enables the model to capture relationships that may not be identified through blood group and location filtering alone.

Different machine learning algorithms use different learning strategies. Logistic Regression provides a probability-based baseline, Decision Tree identifies patterns through feature-based splits, Random Forest combines multiple trees for stable prediction, and K-Nearest Neighbors classifies records based on similarity. Comparing these algorithms under the same conditions helps evaluate their suitability for blood donor availability prediction.

The proposed system connects machine learning prediction with donor identification and record management. When a requester specifies the state, city, and blood group, matching records are retrieved and evaluated using the trained model. Predicted availability prioritizes potentially available donors, reducing manual examination. The system also supports donor registration and periodic information updates. The main contribution of this paper is integrating comparative machine learning evaluation with a functional blood donor search system. Four supervised classification algorithms are evaluated using Accuracy, Precision, Recall, and F1-Score, with the best-performing model integrated into the application. This combines donor matching with availability prediction and provides a data-driven mechanism for prioritizing potentially available donors.

2. EXISTING SYSTEM

Machine learning has been applied in blood-related applications to improve donor management and blood supply planning. Different studies have explored donor availability, donor behaviour, and blood demand using machine learning techniques.

Sivasankaran et al. [1] proposed a machine learning approach for predicting stem cell donor availability. Their work demonstrated that donor-related information could be used to estimate donor availability and support the donor selection process.

Li et al. [2] investigated machine learning methods for predicting the availability of hematopoietic stem cell donors. They compared different machine learning algorithms, including Logistic Regression, Support Vector Machines, and Boosted Decision Trees, and showed that predictive models can assist in identifying donor availability.

Wu et al. [3] used machine learning to predict the willingness of experienced blood donors to donate. Their study considered demographic and donation-related factors to understand donor behaviour and demonstrated the potential of machine learning for improving blood donor recruitment.

Shih and Rajendran [4] compared time-series methods and machine learning algorithms for forecasting blood supply. Their findings showed that predictive models can be used to support blood supply planning and help blood services anticipate future blood requirements.

Ding et al. [5] developed a hybrid neural-network-based model for blood donation forecasting. Their work demonstrated how neural network techniques can be applied to predict blood donation patterns and support effective blood supply management.

These works show that machine learning can be useful in blood donor management and blood supply prediction. However, the present paper focuses on predicting individual donor availability by comparing Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors using donor demographic, medical, location, and donation-history attributes.

3. RESEARCH METHODOLOGY

The proposed system uses supervised machine learning methods to predict blood donor availability. The overall process consists of several stages, including dataset collection, data preprocessing, model training, performance evaluation, model comparison, and model selection. The donor dataset is first prepared by selecting relevant donor attributes and removing unnecessary information. The categorical features are converted into numerical form before training the models. Four supervised machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN), are trained using the preprocessed donor data. Their performance is compared using standard classification metrics, and the best-performing model is selected for the proposed blood donor search system. The selected model is then integrated with location and blood-group based donor searching to predict availability and prioritize matching donors.

3.1 Dataset Description

The proposed paper uses a synthetic donor dataset containing demographic, medical, location, and previous-donation information. The dataset includes both numerical and categorical attributes that can be used to identify patterns related to donor availability. The selected features include age, gender, weight, blood group, city, state, medical condition, hemoglobin level, days since last donation, and previous donation count. Each feature provides different information about the donor that can contribute to availability prediction.

The target variable is **Availability**, which represents whether a donor is available or unavailable. Attributes such as age, gender, and weight provide demographic information, while medical condition and hemoglobin level provide health-related information. Blood group, city, and state provide information required for matching donors with a blood request. Days since last donation and previous donation count represent the donor's previous donation history. Personal identification and contact-related fields such as Donor ID, Donor Name, Contact Number, and Email are not used as prediction features.

3.2 Data Preprocessing

The donor dataset is preprocessed before training the machine learning models. Unnecessary attributes that are not required for availability prediction are removed. Categorical features such as gender, blood group, city, state, and medical condition are converted into numerical representations using Label Encoding. This allows the machine learning algorithms to process both categorical and numerical donor information.

The input features are separated from the target variable, Availability. The prepared dataset is then divided into training and testing data using a 70:30 ratio. The training data is used to learn patterns from existing donor records, while the testing data is used to evaluate the performance of the trained models on unseen records. A random state of 42 is used to maintain consistent data splitting during the experiment.

3.3 Machine Learning Models

Four supervised classification algorithms are used for comparison in this paper. These models represent different approaches to classification, including probability-based learning, tree-based learning, ensemble learning, and distance-based learning. Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors are trained using the same preprocessed dataset to provide a fair comparison of their performance for donor availability prediction.

3.3.1 Logistic Regression

Logistic Regression is a probability-based classification algorithm used to predict whether a donor belongs to the available or unavailable class. It is included as a baseline model because of its simple structure and probability-based output. The sigmoid function is used to convert the input values into a probability:

$$P(y = 1 | X) = 1 / (1 + e^{-z})$$

where,

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Here, (X_1) to (X_n) represent the donor features, (β_0) represents the intercept, and (β_1) to (β_n) represent the model coefficients. The calculated probability is compared with a classification threshold to determine the final availability class.

3.3.2 Decision Tree

Decision Tree is a tree-based classification technique that makes decisions using a sequence of feature-based conditions. The dataset is divided into smaller groups based on feature values until a final classification is obtained. For example, donor records can be separated based on medical condition, hemoglobin level, or donation history to identify different availability patterns.

Gini Impurity can be used to determine the quality of a split:

$$Gini = 1 - \sum_{i=1}^C p_i^2$$

where (C) represents the number of classes and (p_i) represents the proportion of samples belonging to class (i). A lower Gini value indicates a purer group. The tree continues splitting the donor records until a final Available or Unavailable classification is produced. This allows the model to capture non-linear relationships between donor characteristics and availability.

3.3.3 Random Forest

Random Forest is an ensemble learning method that combines multiple Decision Trees to produce a final prediction. Each tree learns from different training samples and feature combinations. The predictions from the individual trees are combined using majority voting.

The final prediction can be represented as:

$$\hat{y} = \text{mode}\{T_1(X), T_2(X), \dots, T_B(X)\}$$

where (B) represents the total number of trees and ($T_1, T_2, \dots, T_B$) represent the individual Decision Trees. The final class is determined by the majority of the tree predictions.

Random Forest is used in this paper because combining multiple trees can provide more stable predictions than relying on a single Decision Tree. It can also identify different patterns among donor demographic, medical, location, and donation-history features. Random Forest is considered as a candidate model for the proposed system, and its final selection is based on the comparative evaluation results.

3.3.4 K-Nearest Neighbors

K-Nearest Neighbors is a distance-based classification algorithm that predicts the class of a new record by comparing it with existing records in the training dataset. The algorithm identifies the nearest donor records according to their feature values and assigns the class that is most common among the selected neighbors.

Euclidean distance can be calculated as:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

where ($d(x, y)$) represents the distance between two records, (x_i) represents the feature value of the new donor record, (y_i) represents the corresponding feature value of an existing donor record, and (n) represents the number of features. Donors with smaller distances are considered more similar. The availability class of the nearest records is then used to classify the new donor.

3.4 Performance Comparison of Machine Learning Models

The four machine learning models are evaluated using the same testing dataset. Their performance is compared using Accuracy, Precision, Recall, and F1-Score. The experimental results are presented in Table 1.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	81.17%	77.51%	99.74%	87.23%
Decision Tree	86.50%	82.69%	100%	90.53%
Random Forest	94.33%	91.92%	100%	95.79%
K-Nearest Neighbors	67.67%	67.97%	94.32%	79.00%

Table 1. Performance Comparison of Machine Learning Models

The results show differences in the performance of the four algorithms. Logistic Regression provides a baseline level of performance, while Decision Tree improves the classification results by learning feature-based relationships. Random Forest achieves the highest Accuracy, Precision, and F1-Score while also achieving 100% Recall. KNN provides comparatively lower performance for the given dataset. Based on the overall experimental results, Random Forest is selected as the prediction model for the proposed donor availability system.

3.5 Model Evaluation Metrics

Accuracy, Precision, Recall, and F1-Score are used to evaluate the four machine learning models. These metrics measure different aspects of the model's ability to classify donors as Available or Unavailable. Using multiple evaluation metrics provides a more complete comparison than relying only on Accuracy.

3.5.1 Accuracy

Accuracy represents the percentage of donor records correctly classified by the model out of all tested records. It considers both correctly predicted Available and Unavailable records.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

where TP represents correctly predicted available donors, TN represents correctly predicted unavailable donors, FP represents unavailable donors incorrectly predicted as available, and FN represents available donors incorrectly predicted as unavailable. A higher Accuracy indicates that a larger proportion of donor records are correctly classified.

where,

TP $\Rightarrow$ True Positive

TN $\Rightarrow$ True Negative

FP $\Rightarrow$ False Positive

FN $\Rightarrow$ False Negative

3.5.2 Precision

Precision represents the proportion of donors predicted as Available that are actually Available. It indicates how reliable the model's positive availability predictions are.

$$\text{Precision} = TP / (TP + FP)$$

A higher Precision means that fewer unavailable donors are incorrectly identified as available.

3.5.3 Recall

Recall measures the ability of the model to correctly identify donors who are actually Available. It is particularly important in donor availability prediction because missing an available donor may reduce the number of potentially useful donor records identified by the system.

$$\text{Recall} = TP / (TP + FN)$$

A higher Recall indicates that the model identifies a greater proportion of the actual available donors.

3.5.4 F1-Score

F1-Score combines Precision and Recall into a single measure. It is useful for comparing models when both reliable positive predictions and identification of available donors are important.

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

A higher F1-Score indicates a better balance between Precision and Recall. Therefore, F1-Score is considered along with Accuracy, Precision, and Recall when comparing the four models.

4. RESULTS AND DISCUSSION

The four machine learning models, Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN), were evaluated using Accuracy, Precision, Recall, and F1-Score. The models were trained and tested using the same preprocessed donor dataset. The numerical results are presented in Table 1, while Figure 1 provides a graphical comparison of the model performances.

The results show clear differences among the four algorithms. Logistic Regression achieved 81.17% Accuracy and 87.23% F1-Score, while Decision Tree achieved 86.50% Accuracy and 90.53% F1-Score. Random Forest achieved the highest Accuracy of 94.33%, Precision of 91.92%, Recall of 100%, and F1-Score of 95.79%. KNN achieved 67.67% Accuracy and 79.00% F1-Score.

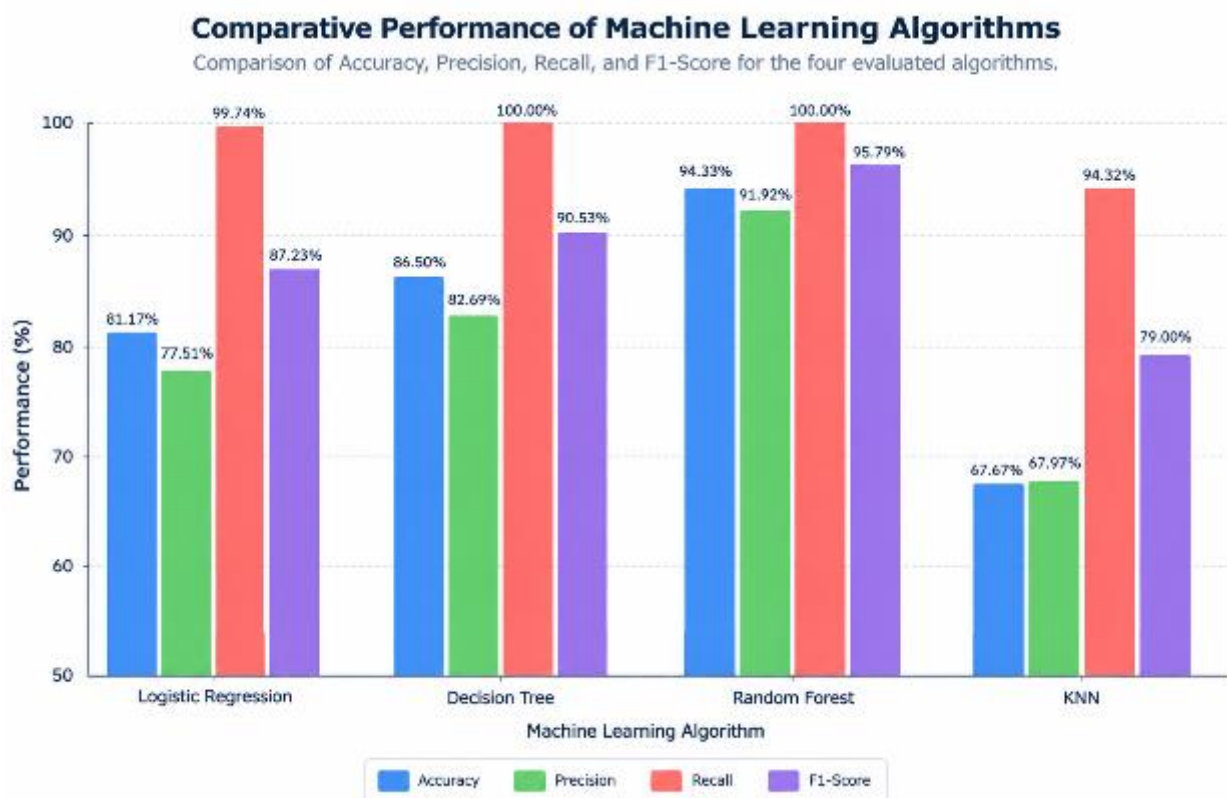


Figure 1. Performance Comparison of Machine Learning Models

Random Forest achieved the best overall performance among the four machine learning models evaluated. Its 100% Recall indicates that all available donor records in the testing dataset were correctly identified. The Precision of 91.92% shows that most donors predicted as available were actually available. In addition, the F1-Score of 95.79% indicates a good balance between Precision and Recall.

The feature importance analysis showed that Medical Condition was the most influential factor in predicting donor availability, with an importance value of 0.5547. Hemoglobin Level was the second most important feature, with a value of 0.3381, followed by Weight (0.0616) and Gender (0.0237). Other features, including Age, Blood Group, City, State, and Previous Donation Count, had comparatively lower importance in the prediction process.

The developed system uses Firebase as its backend database to store and manage donor information. New donor profiles can be registered, and existing profiles can be updated whenever necessary. When a user searches for a donor, the system retrieves matching records based on the selected blood group, state, and city. The Random Forest model then predicts donor availability, and the results are arranged to give higher priority to donors with a greater predicted likelihood of being available.

Based on the comparative evaluation, Random Forest was selected as the final model and integrated into the donor search system. By combining machine learning predictions with blood group and location-based filtering, the system helps users identify suitable donors more efficiently. It also organizes the search results according to predicted availability, making it easier to find potentially available donors.

Overall, the results highlight the usefulness of machine learning in predicting blood donor availability. The developed system combines donor information management, availability prediction, and donor search in a single application. It can help simplify the process of finding potential blood donors during emergencies, although actual donor availability must still be confirmed before donation.

5. CONCLUSION

In this paper, four machine learning algorithms, Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN), were evaluated for predicting blood donor availability. The four models were evaluated and compared based on their Accuracy, Precision, Recall, and F1-Score.

Among the four models, Random Forest achieved the highest Accuracy of 94.33%, Precision of 91.92%, and F1-Score of 95.79%, with 100% Recall. Based on these results, Random Forest was selected as the final prediction model for the proposed blood donor availability system.

The analysis revealed that Medical Condition and Hemoglobin Level played an important role in predicting donor availability. Other donor attributes such as Weight, Gender, Age, Blood Group, City, State, and Previous Donation Count also contributed to the model.

The selected model was integrated with the donor search application. The system uses State, City, and Blood Group to identify matching donors and predicts their availability. The predicted availability is then used to prioritize potentially available donors.

The system can be improved in the future by using updated real-world donor data, real-time donor availability, additional machine learning algorithms, and advanced donor prioritization methods. A real-time notification mechanism can also be incorporated to notify suitable donors when their blood group and location match an emergency request. Overall, this paper demonstrates how machine learning can support faster and data-driven blood donor identification.

REFERENCES

- [1] A. Sivasankaran, E. Williams, M. Albrecht, G. E. Switzer, V. S. Cherkassky, and M. Maier, "Machine Learning Approach to Predicting Stem Cell Donor Availability," *Biology of Blood and Marrow Transplantation*, vol. 24, no. 12, pp. 2425–2432, 2018. DOI: 10.1016/j.bbmt.2018.07.035.
- [2] Y. Li, A. Masiliune, D. Winstone, L. Gasieniec, P. Wong, H. Lin, R. Pawson, G. Parkes, and A. Hadley, "Predicting the Availability of Hematopoietic Stem Cell Donors Using Machine Learning," *Biology of Blood and Marrow Transplantation*, vol. 26, no. 8, pp. 1406–1413, 2020. DOI: 10.1016/j.bbmt.2020.03.026.
- [3] H. Y. Wu, Z. G. Li, X. K. Sun, W. M. Bai, A. D. Wang, Y. C. Ma, R. H. Diao, E. Y. Fan, F. Zhao, Y. Q. Liu, et al., "Predicting Willingness to Donate Blood Based on Machine Learning: Two Blood Donor Recruitments During COVID-19 Outbreaks," *Scientific Reports*, vol. 12, p. 19165, 2022. DOI: 10.1038/s41598-022-21215-2.
- [4] H. Shih and S. Rajendran, "Comparison of Time Series Methods and Machine Learning Algorithms for Forecasting Taiwan Blood Services Foundation's Blood Supply," *Journal of Healthcare Engineering*, vol. 2019, Article ID 6123745, pp. 1–6, 2019. DOI: 10.1155/2019/6123745.
- [5] X. Ding, X. Zhang, X. Li, and J. Du, "A Hybrid Neural Network Based Model for Blood Donation Forecasting," *Journal of Biomedical Informatics*, vol. 146, p. 104488, 2023. DOI: 10.1016/j.jbi.2023.104488.