A survey on machine learning algorithms for the blood donation supply chain

Subramanian Mahadevan1, S Poornima2, Kartikeya Tripathi3, M Pushpalatha4

1,3,4th Year/CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

2Assistant Professor/CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

4Professor/CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

sma
devan_ms@srmuniv.edu.in1, poornima.se@ktr.srmuniv.ac.in2, kartikeyat
tripathi_sa@srmuniv.edu.in3, pushpalatha.m@ktr.srmuniv.ac.in4

Abstract- With the proliferation of big data, the need for intelligent and automated systems has risen. This need is probably felt the most in the field of health care, especially in the area of blood transfusion, since they require supplies at the earliest. Currently, transfusion services are heavily manual in nature, which is not ideal. The rising demand for blood and the decline in donation rates has put a lot of strain on the blood donation supply chain. Hence, creating intelligent systems that can make decisions and improve communication across the supply chain is of great importance. In this paper, we are going to give a general summary of the various machine learning techniques which have been applied to this domain and compare their advantages and disadvantages.

1. INTRODUCTION

A blood transfusion may be scheduled or be needed urgently. However, blood cannot be manufactured artificially; but it can be given as a gift to someone who direly needs it. This humanitarian act might provide the receivers some more time to spend with their loved ones. Blood distribution is an important activity within the blood supply chain as allocation of the donated blood lives every day. Unfortunately, in 2016-17, India faced a deficiency of 1.9 million units of blood, which is analogous to 60 tankers. This in part can be attributed to the lack of proper communication between blood banks, hospitals and genuine donors who coincidentally might have been present in greater proximity at the time of need. An analysis on blood donation trends after a temporary deferral indicates the amount of factors that influences a person before returning to donate blood.

Due to these problems faced by transfusion services, the need for an intelligent automated agent is of paramount importance. The incorporation of information technology in this domain can help in decreasing the workload and identify donors on the basis of their past medical conditions. In this paper, we shall compare and contrast the different machine learning algorithms which have been applied to the domain of medicine and assess whether or not they are a viable candidate for improving the blood donation supply chain.
2. ALGORITHMS

A. Logistic Regression

Developed by David Cox in 1958, logistic regression is a popular method to solve binary and multivariate classification problems. As the name suggests, it’s named after the logistic function.

\[ f(x) = \frac{1}{1 + e^{-x}} \]

Referred to as the sigmoid function, it accepts any real valued number and outputs a numeric value between 0 and 1. If the value is above a certain numeric threshold (normally taken as 0.5 for the purpose of calculating a probability), a TRUE value is given as the outcome. If it is below the threshold, a FALSE value is shown. The cost function used to modify the value of the predicted output is given as follows:-

\[ cost(f(x), y) = \begin{cases}  
-\log(f(x)), & \text{if } y = 1 \\
-\log(1 - f(x)), & \text{if } y = 0 
\end{cases} \]

Logistic Regression has various applications in the medical field such as predicting severity of a patient, risk of contracting a disease based on available factors and mortality rate. As an example, The TRISS (Trauma and Injury Severity Score) was created by Boyde CR using this algorithm. Some of the other applications for this algorithm include predicting the voting patterns of the general public, finding out the probability of failure of a product and speculating about mortgage defaulters.

B. Artificial Neural Networks

Artificial neural networks are based on the functioning of biological brains, which consist of an interconnected set of neurons. The first layer is called the input layer, which feeds the initial input to the model, and the final layer is the output layer, which produces the terminal output. Between those two, there may be one or more hidden layers. Each neuron at one level is connected to every other neuron in the next level. On receiving an input, a neuron applies an activation function to the signal, which introduces nonlinearity to the model. A weight matrix, which is applied to the input signals between layers, is modified to train the network, using backpropagation. The given figure demonstrates the rudimentary working of a typical artificial neural network.
Feedforward Neural Network is one of the simplest implementation of an ANN. This network has its input signal being transmitted in a single direction, from the input layer to the output layer. Based on the kind of network being used (single or multi-layer perceptron), a different activation function would be used. The logistic or sigmoid function is a common choice for the activation function of a single layer perceptron.

### C. Decision Trees

Decision Trees cover both classification and regression problems. It uses a tree structure to evaluate decisions and their ramifications, while considering utility costs. It’s primarily used in decision analysis problems, to identify a path most probable to reach the favorable outcome. It is typically represented as a flowchart with the objective of highlighting all the possible outcomes to a given decision.

![A generic decision tree structure](image)

**Fig. 2:** A generic decision tree structure
Three types of nodes are present in decision trees:-

1. Decision nodes — the point at which the flow of the tree is evaluated against a condition and based on the response, branches are formed.

2. Chance nodes — Used to represent the different outcomes branching from a decision node. It does not evaluate the flow of the tree with a condition, but used to track the different factors involved in a decision. Ex. In the above figure, high and normal describes the state of humidity.

3. End nodes — these are the final terminating nodes of the tree, and don’t have any additional child nodes. They represent the final states of a decision tree.

Decision trees are of two types:-

1. Classification trees
2. Regression trees

Some of the popular algorithms for decision trees are CHAID or Chi squared automatic interaction detector, C4.5, CART or Classification and Regression Tree, MARS and ID3.

Developed by Brieman in 1984, CART algorithm is one of the more popular decision tree algorithms and can be used to create both classification and regression trees. Classification trees are constructed by the binomial splitting of a feature, using Gini Index to select the splitting attribute. This index only performs binary splits.

Decision trees are easy to construct and understand. They help in providing insight about situations with hard data. Due to the probabilities, branches and utility costs associated with a tree, they provide an expert description of a problem statement. But they are also very expensive to create, particularly for datasets with a large number of outcomes and uncertainty in values.

D. Support Vector Machines

A supervised learning algorithm having uses in both classification and regression problems, SVM’s are formally defined by a hyperplane which separates all the data points in an N-dimensional space. When trained with a set of examples, an SVM builds a non probabilistic binary linear classifier. To separate the data points in space, multiple hyperplanes are possible choices for an SVM. The optimal hyperplane is chosen by maximizing the distance between the data points of both classes. This is done so that future data points can be mapped more accurately. The number of features of the model determines the dimension of the hyperplane. Logistic regression uses the sigmoid function to squash the output and limit it to the range [0, 1]. If the squashed value is less than or greater than a predefined threshold, the appropriate class label is assigned. Unlike logistic regression, SVM takes the output of a linear function and checks whether the value is in the range [-1, 1]. This range acts as the margin to be maximized between data points.
To maximize the margin between data points and the, the following loss function is used:

\[ c(x, y, f(x)) = (1 - y \cdot f(x))^+ \]

A regularization parameter is used to balance the margin maximization and loss. With regularization parameter, the loss function is given by:

\[ \min_{\omega} \frac{1}{2} \| \omega \|^2 + \sum_{i=1}^{n} (1 - y_i (x_i , \omega))^+ \]

3. LITERATURE SURVEY

To predict the likelihood of an individual to donate blood certain factors have to be taken into account. Factors like gender, ethnicity religion, civil status, income level, age, employment, distance problem, lack of direction, self-interest, fear of hospital setup and recency of donation. The relevant and respective data taken from all fields when fed into suitably structured logistic regression model generates a donor score [1], which can give a rough indication of the probability of an individual to donate blood. But due to the lack of a robust dataset to train and overfitting the accuracy of the logistic regression model is not very reliable.

Alternative supervised learning models can also be explored for donor and non-donor classification based on arbitrary factors. Artificial Neural Networks and Decision trees provide comparable and better performing models respectively. The results they generate are closely comparable. A Multilayer Perceptron with backpropagation was used for the accuracy test of donor group classification. The model classified candidates into the blood donor group with recall and precision values of 88% and 81.7%, respectively, [2] and classified the others into the non donor group with recall and precision values of 41.6% and 53.8% respectively. The total accuracy of this model was 76.25%. ANN indicated to be a more suitable model for such computations.

Since the blood donation system is also a supply chain there are possibilities of optimizing it by using machine learning techniques like artificial neural networks. Constructing a supply chain framework with hybrid inventory decisions of supply chain numbers with the consideration of factors that impact the total supply chain cost [3], provides a more streamlined approach which influences Optimal inventory policies that are derived by using multilayer perceptrons and backpropagation Artificial Neural Networks.
Network models.

Solving a humanitarian problem should not create more of them. Software sustainability is of great importance with high number of smartphone users in today’s day and age. Assessment of greenability of Blood Donation apps with respect to all aspects is an issue which should be taken into consideration. It would help plan and formulate a catalog of reusable requirements in order to develop a green Blood Donation app [4], which may be refined, expressed and adapted in the form of system and software requirements. However, an exclusion of the economic dimension is overlooked which should be worked on and included.

An alternative to the existing blood donation conundrum is the generic electronic donor card system which makes use of DBMS and Electronic data exchange [5]. The highlight of introducing and implementing such a possibility is overcoming the human-machine interaction challenges and physical redundancy like filling forms along with labour intensive tasks like manual updating. The blood donation centre staff faces difficulties in using such an unfamiliar system. Project Management Unit or PMU need to be positioned at different centers to take care of specialized issues [6]. The PMU also needs to teach the staff to resolve technical issues, due to the demanding nature of the job.

Attempts to combat such limitations look to automation using another alternative which includes new hardware entities like a Raspberry Pi based app system [7]. Registering of new users and medical requests can be handled using such an android app, with the user details being stored in a database. When formulating a request, the user specifies the blood group and current address. Based on the request, corresponding donor information from the database is displayed and an SMS is sent to all the donors using a GSM modem. But in the absence of a methodology that identifies optimal donors and the tedious set up of physical systems at various locations, it is not a viable solution.

Based on what we have learnt until now, it seems natural to compare ANN and LR models [8] . When comparing the two algorithms on a clinical dataset of hepatocellular carcinoma surgery patients, sex, age, survival percentage of hospital, hospital traffic and surgeon availability were some of the patient features that were analyzed. The logistic regression model had a superior AUROC [9] in 84.67% of the trials, better accuracy rate in 97.28% of the experiments, a lower Hosmer-Lemeshow statistic [10] in 41.18% of the cases. Nevertheless, it is hard to ignore the paper’s lack of consideration for the complications associated with HCC surgical operations as a viable factor.

It is essential to find out factors which influence people’s decision when donating blood, especially after a deferral. When semi structured interviews of regular blood donors following a short term deferral were conducted, individuals expressed their improvement in self health concepts and ease with which they could help a fellow being as chief reasons for donation [11] After a deferral, their return depended on mainly four factors: standing obligations, support system for donation, feeling of self-worth and the presence of a strong blood donor identity. But as one can surmise, the reliability of the test subjects is under question due to them not being vetted properly.

A method to calculate the donor’s probability of donating within a specified time period is described with a probability function. \( P(D_{ts-te}) \) is calculated based on the donation history of the subject, where "ts" and "te" represent the time interval [12]. A donor score is devised for both first time and repeat donors. A logistic regression model is created, which models the probability of donation by applying two variables — a score for repeat donors and a variable representing a first time donor. This method has the limitation of lack of generalization.
The shelf life of blood and its constituents is quite low. Therefore regular and synchronized blood drives have to be conducted. Mobile blood drives are one of the more popular ones. A two-step framework can be implemented to plan the mobile blood collection. A mathematical model is proposed to compute the collections of each day and a probabilistic model is utilized to make donation forecasts [13]. Use of such operational and tactical models makes the system more efficient. However, the collection system is only deployed by EFS, France; the conditions and results produced in a single first world country may not necessarily be relevant globally.

Saving a donee’s life is essential but not at the cost of jeopardizing the donor’s life. Data processing can be widely used for classification of people who can safely donate blood and people who have allergic, vasovagal or deficiency-based reactions. Pattern Matching, Clustering Algorithms and Text Mining techniques [14] can be implemented, so we can take timely steps and actions to prevent such harm and facilitate a safe blood donation environment. Analysis of such classified clusters would provide us better insights and make the blood donation process more proficient.

Introduction of data warehousing can be seen as a significant component of the integrated framework to store past blood donation data in a central database for analysis. Proposition of a philanthropy score [15] based on the population’s tendency to do good deeds is controversial. There is also proposed inclusion of Unique Identification/Social Security or their equivalent which would streamline the system to an extent.

There is a wide focus on comparison of Service Vector Machines and Artificial Neural Networks. Multiple studies come out providing more generalizing views on which algorithm does what better. These two algorithms when used to predict currency exchange rate performance respectively provide supplemental result. A variant of SVM called Least squares SVM is used [16]. It is seen that the exchange rate prediction method, based on LS SVM, provides results that aligns adequately with the real exchange rate change, thus overcoming non linearity and randomness in exchange rate. For a small dataset which is short term in nature, the accuracy of the LS SVM implementation is better than that of the ANN.
| Paper | Techniques Used | Disadvantages |
|-------|-----------------|---------------|
| 1     | Logistic Regression | Data is specific to a location and is not in-house. Lack of research on blood donor behavior and characteristics. Bad accuracy of donor score overfitted model. |
| 2     | Back Propagation Artificial Neural Networks | Back Propagation Multilayer perceptron has very poor generalization ability for statistically neutral problems. To find the ideal parameters, trial & error method is employed. |
| 3     | Temperature Tag Barcode Interface SMS and e-Mail Server Biometrics Smart Card | The staffs at the blood donation centers are unfamiliar with such a framework. Project Management Unit or PMU should be positioned at different centers to take care of specialized issues. The PMU also needs to teach the staff to resolve technical issues, due to the demanding nature of service. |
| 4     | Artificial Neural Networks Logistic Regression | Clinical image acquired in this analysis is not as accurate as that of an investigation into clinical trial data due to possible inconsistencies in the formulation of surgical modalities and primary diagnoses. Issues related to the surgical procedures were not evaluated, which limits the plausibility of the comparison. |
| 5     | Artificial Neural Networks Decision Trees | Different components influencing blood donation should be studied to oversee donor behavior. By expanding the sample size of the study, it’s possible to collect more data |
| 6     | Artificial Neural Networks | The dimensionality of the input space has no relation with the computation |
ANN and LS-SVM Exchange Rate Performance Prediction” [16] | Least Squares Support Vector Machine | complexities of SVM’s ANN’s overfit comparatively easily

| “Mathematical Programming Models for Annual and Weekly Bloodmobile Collection Planning” [13] | Mixed Integer programming Models Vehicle Routing Problem | Data evaluated and fed is specific to a single first world country and might be globally irrelevant. If the donation parameters and annual demand are not well known then such planning can prove to be hazardous.

Table 1: Literature Survey Table

4. CONCLUSION
We have presented a brief survey of papers that have done work related to healthcare automation, particularly transfusion medication. We have concluded that machine learning techniques like ANN, SVM and decision trees are vital for the development of a balanced and dependable mechanism for communicating with and identifying viable blood donors. The existing system has physical redundancies which would be solved by integrating machine learning. Initiatives like these would lead to a decrease in response time and avoidable costs would be cut down. The possibility of finding an ideal donor would increase resulting in lower rates of transfusion related complexities.

5. REFERENCES
[1] Raj A, Gupta A & Selvaraj P. 2018. “Predicting Donor’s Likelihood of donating blood given various factors”. International Journal of Pure and Applied Mathematics.118.
[2] Boonyanusith W, Jittamai P. “Blood donor classification using neural network and decision tree techniques”. In Proceedings of the World Congress on Engineering and Computer Science 2012 (Vol. 1, pp. 499-503).
[3] Chen HC, Wee HM, Hsieh YH. Optimal supply chain inventory decision using artificial neural network. In 2009 WRI Global Congress on Intelligent Systems 2009 May 19 (Vol. 4, pp. 130-134). IEEE.
[4] Ouhbi S, Fernández-Alemán JL, Idri A, Pozo JR. Are mobile blood donation applications green?. In 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA) 2015 Oct 20 (pp. 1-6). IEEE.
[5] Li BN, Dong MC. Banking on blood [electronic donor card system]. Computing and Control Engineering. 2006 Aug 1;17(4):22-5.
[6] Cheema AS, Srivastava S, Srivastava PK, Murthy BK. A standard compliant blood bank management system with enforcing mechanism. In 2015 International Conference on Computing, Communication and Security (ICCCS) 2015 Dec 4 (pp. 1-7). IEEE.
[7] A. C. Adsul, V. K. Bhosale and R. M. Autee, "Automated blood bank system using
Raspberry Pi," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 252-255.

[8] Shi HY, Lee KT, Wang JJ, Sun DP, Lee HH, Chiu CC. Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma: a nationwide study. Journal of gastrointestinal surgery. 2012 Nov 1; 16(11):2126-31.

[9] Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine. 2013; 4(2):627.

[10] Lemeshow S, Hosmer Jr DW. A review of goodness of fit statistics for use in the development of logistic regression models. American journal of epidemiology. 1982 Jan 1; 115(1):92-106.

[11] Hillgrove TL, Doherty KV, Moore VM. Understanding non-return after a temporary deferral from giving blood: a qualitative study. BMC public health. 2012 Dec; 12(1):1063.

[12] Flegel WA, Besenfelder W, Wagner FF. Predicting a donor's likelihood of donating within a preselected time interval. Transfusion Medicine. 2000 Sep 1; 10(3):181-92.

[13] Alfonso E, Augusto V, Xie X. Mathematical programming models for annual and weekly bloodmobile collection planning. IEEE Transactions on Automation Science and Engineering. 2015 Jan; 12(1):96-105.

[14] Ghodekar P, Wagh S, Fulsundar S. Reduce Complexity of Blood Donation Process And Make It Safe by Using Data Mining. International Journal of Advance Engineering and Research Development. December 2016. 3(12). 2348-4470.

[15] Maji G, Debnath NC, Sen S. Data Warehouse Based Analysis with Integrated Blood Donation Management System. In 2018 IEEE 16th International Conference on Industrial Informatics (INDIN) 2018 Jul 18 (pp. 855-860). IEEE.

[16] Sun A, Zhao T, Chen J, Chang J. Comparative Study: Common ANN and LS-SVM Exchange Rate Performance Prediction. Chinese Journal of Electronics. 2018 May 1; 27(3):561-4.

[17] NHK K. ISMAIL*,”Estimation Of Reliability Of D Flip-Flops Using Mc Analysis”, Journal of VLSI Circuits And Systems 1 (01), 10-12,2019.

[18] MN BORHAN,"Design Of The High Speed And Reliable Source Coupled Logic Multiplexer”,Journal of VLSI Circuits And Systems 1 (01), 18-22,2019