E-ISSN: 2676-6167 | P-ISSN: 2538-5100

Journal of Applied Research on Industrial Engineering

www.journal-aprie.com

J. Appl. Res. Ind. Eng. Vol. 10, No. 2 (2023) 218–237.

Paper Type: Research Paper

# A New Method to Predict the Quality of Umbilical Cord Blood Units based on Maternal and Neonatal Factors and Collection Techniques

**Rasoul Jamshidi[1,*], Sattar Rajabpour Sanati[2], Morteza Zarrabi[3]**

[1] Department Industrial of Engineering, School of Engineering, Damghan University, Damghan, Iran; r.jamshidi@du.ac.ir.

[2] Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran; sanati.sattar@gmail.com.

[3] Royan Institute, Royan Stem Cell Technology Company, Tehran, Iran; m.zarrabi@rsct.ir.

**Citation:**

Jamshidi, R., Rajabpour Sanati, S., & Zarrabi, M. (2023). A new method to predict the quality of umbilical cord blood units based on maternal and neonatal factors and collection techniques. *Journal of applied research on industrial engineering, 10*(2), 218-237.

Received: 14/05/2021 Reviewed: 15/06/2021 Revised: 11/08/2021 Accept: 20/09/2021

## Abstract

The saving banks of "umbilical cord blood stem cells" are considered as strategic health-based institutions in most countries. Due to the limited capacity of cord blood sample storage tanks, the samples should be evaluated according to their quality. So these banks need a method to assess quality. In this paper, first, the effective factors on the quality index of the extracted cord blood from newborn infants are identified using the electronic records and database of Royan's umbilical cord blood bank. Then by machine learning and various statistical methods such as Multilayer Perceptron Neural Networks (MLPNNs), Radial Basis Function Neural Networks (RBFNNs), Logistic Regression (LR), and C4.5 Decision Tree (DT), the quality value of blood samples and their proper category (for discarding or freezing) are determined. Two different sets of data have been used to evaluate the proposed methods. The results show that the ensemble of RBFNN with k-means clustering model has the best accuracy compared to other methods, which categorizing the samples with 91.5% accuracy for the first data set and 81.6% accuracy for the second one. The results also show that using this method can save about 1 million dollars annually.

**Keywords:** Umbilical cord blood banking, Data mining, Neural network.

## 1 | Introduction

The stem cells are unique for their high proliferation ability and convertibility to other different cell types. The blood of umbilical cord is a precious resource of the stem cells, and it is a good alternative for bone marrow transplantation. Today, about 20 percent of stem cell transplantations are originated from the stem cells of the umbilical cord blood. Transplantation of the stem cells has various advantages such as facile collection and access, being safe and risk-free for mothers and newborns, lack of the need to total Human Leukocyte Antigen compatibility, reduction of the likelihood of transplant rejection, and Graft-Versus-Host Disease reaction. Despite these numerous advantages, using the umbilical cord blood has some limitations such as low volume, slow deployment process of the transplant in the host's body, and delays in the neutrophilic and platelet



Corresponding Author: r.jamshidi@du.ac.ir

https://doi.org/10.22105/jarie.2021.286186.1326

cell retrieving recovery. Therefore, there is a vital need for transplantable stem cell detection in the umbilical cord blood and their safe storage [1]-[3].

Actually, storage of the umbilical cord blood cells provides an opportunity for a ready to use and always available resource of the stem cells, which is genetically compatible with the donor and it can be used to cure the possible future diseases of the individual or even his/her family members [3], [4].

Storage of the umbilical cord blood samples is carried out in three types of banks: public or national banks, private or family banks, and hybrid umbilical cord blood banks. The growth rate of these banks by 2011 based on their types is illustrated in *Table 1*. The growth rate of private banks has been significantly higher than the public ones.

**Table 1. The worldwide growth rate of the umbilical cord blood banks.**

| Number | Year 2003 | Year 2004 | Year 2011 | The Growth |
|---|---|---|---|---|
| Public banks | 18 | 20 | 29 | 61% |
| Private banks | 15 | 21 | 30 | 100% |
| Samples in the public banks | 58650 | 76145 | 160000 | 2.7 times |
| Samples in private banks | 179350 | 270991 | 1150000 | 6.4 times |

If an accurate evaluative prediction of the future samples is provided at any stage before the conclusion of the contract (between the banks and families) or before conducting the qualitative tests, the umbilical cord blood banks can avoid possibly useless tests and cost. Also, due to the limited capacity of the storage tanks of the umbilical cord blood samples, the banks should prioritize their blood samples based on quality factors. If the samples are not ideal in terms of their quality level, they cannot be used at the time of transplantation. That is why the storage of high-quality samples has vital importance. Prioritization of samples should be based on parameters extracted from historical data by data mining and machine learning methods [5]-[7].

In this paper, the data mining techniques are used for the first time to predict the quality of the umbilical cord blood samples. Based on this prediction, sample quality is evaluated, and a decision is made about sample storage. Implementation of the proposed method reduces storage costs and also increases the likelihood of using these samples to treat diseases in the future.

The previous studies mainly focused on the factors affecting the quality of umbilical cord blood and proposed some methods for cell counting and samples selection, but in the proposed method, we predict the samples' quality to select the best samples for storage. To predict the sample quality, some heuristic methods such as Multilayer Perceptron Neural Networks (MLPNNs), Radial Basis Function Neural Networks (RBFNNs), Logistic Regression (LR), and C4.5 Decision Tree (DT) are used. Also, a case study is presented at Royan institute, which shows the efficiency of the proposed method.

## 2 | Literature Review

Many studies have been dedicated to evaluate the factors affecting the quality of blood samples extracted from the umbilical cord. Al-Sweedan et al. [8] determined the effective factors on the number of the hematopoietic stem cells collected from the umbilical cord blood. The data of 200 individuals which were eligible for blood-producing tests-such as Total Nucleated Cells (TNCs) and the number of CD 34+ cells, gathered for analysis. The results have been evaluated by single and multivariate analysis. In the single variable analysis, the factors with a positive correlation to TNC numbers were: maternal weight, preeclampsia, neonatal weight, neonatal platelet count, neonatal Rh, gestational age, and delivery type. Also, the positive factors related to the high number of CD34 + cells were: maternal weight, preeclampsia, maternal hypertension, neonatal weight, neonatal Rh type, and delivery type [8].

Lee et al. [9] analyzed the effective intrinsic factors on the hematopoietic variables of cord blood in Korea's newborns. The total number of nucleated cells, CD34+ cells, and the ratio of CD34+ cells to the whole

nucleated cells in infants were compared regarding some factors such as sex, gestational age, birth weight, birth weight centile for gestational age, and blood type (ABO). The results provided that the TNC number was lower in males, but there was no difference in the CD34+ number. The increase in gestational age was positively correlated with whole nucleated cells, while it led to a decrease in the number of CD34 + cells and reduced the rate of CD34+ cells to the entire nucleated cells. The TNCs, CD34 + cells, and the ratio of CD34+ to TNCs increased when the neonatal weight increased. Abdul Wahid et al. [10] evaluated a comparison between the number of hematopoietic stem cells and CD34+ of the umbilical cord blood in preeclampsia, and those of the control sample and the factors affecting these observations. The cord blood volume, nucleated count, and CD34+ cell count in PE subjects were significantly lower than the non-PE subjects. Jan et al. [11] discussed whether maternal factors such as age, race, and ethnicity affect the laboratory parameters of hematopoietic content, such as CD34+ cells count, TNCs umbilical cord blood volume. The effect of neonatal factors, including the delivery type, gestational age, gender, and birth weight on those parameters was also studied as well. Ballen et al. [12] studied the effects of various factors such as race, age, and smoking habits on the potential laboratory parameters of hematology. Furthermore, the neonatal characteristics such as birth order, neonatal weight, gender, and gestational age on laboratory factors were analyzed. Surbek et al. [13] studied the effect of the extraction time of the umbilical cord blood. They argued that collecting the cord blood before or after the outgoing of the placenta may affect the volume of collected blood. Al-Deghaither [14] analyzed maternal and neonatal factors, including maternal age, neonatal weight, placental weight, neonatal gender, and the number of pregnancies. They showed that if the neonatal weight is more than 3.3 kg, the gestational age is lower, the size of placenta is larger, and the baby is delivered in the first or second pregnancy, TNC, CD34+, CD45+, NRBC, and viability will be higher. Nunes and Zandavalli [15] studied the effects of maternal and fetal factors on the qualitative characteristics of the umbilical cord blood in public blood banks. In this study, the factors of gestational age, mode of delivery, and neonatal weight of 458 samples were analyzed. The qualitative factors of blood samples, TNC, CD34+, and blood volume were considered. Manegold-Brauer et al. [16] investigated maternal factors such as maternal age, number of pregnancies, pre-gestational weight, as well as neonatal factors such as neonatal gender and birth weight. The focus was on the prediction of TNC index, which is just one of the various qualitative characteristics of the umbilical cord blood sample. Donaldson et al. [17] used the data of 500 umbilical cord blood samples to study the effects of factors such as maternal age, gestational age, neonatal weight, placental weight, duration of the first test, duration of the second test, total duration of the tests, duration of the collecting process of the placental blood, and duration of the collecting process of the venous blood of the umbilical cord on the qualitative indices such as volume and TNC.

Also, some researchers investigated the methods implemented for counting the cells in the blood unit. Jaime-Pérez et al. [18] analyzed the current standard method using the volume and TNC count to select cord blood units, cryopreservation, and further transplantation. The data consisted of 794 units of umbilical cord blood, which contained CD 34+ cells determined by flow cytometry. Wen et al. [19] investigated the relationship between the factors associated with the donor and the umbilical cord blood quality indices. The obstetric and neonatal clinical laboratory data of 1549 units of umbilical cord blood were gathered from the Buddhist Tzu Chi Stem Cells Center. A multivariate analysis method was used to analyze the data. The results showed that the neonatal birth weight has a significant positive correlation with any of the clinical features, i.e., the number of CD34+ cells, TNC count, unit volume, and placental weight. Cobellis et al. [20] examined the question of whether the storage of the umbilical cord blood using ultrasound and sonographic parameters at the time of pregnancy is predictable. For this purpose, the correlation of all sonographic parameters (head width, head circumference, abdominal circumference, femur length, estimation of fetal weight, rate index of the umbilical artery), which were extracted from the newborn's weight at birth and placental weight, were studied with the storage parameters of the umbilical cord blood samples (volume, CBU, TNC, and CD34+). After analyzing the 219 pregnant women, the results suggested that some factors such as abdominal circumference, femur length, estimation of fetal weight, neonatal weight at birth, and placental weight have a positive effect on the storage parameters such as CBU, CBU volume, TNC, and CD34+.

Optimizing cells selection to reduce the storage cost and the high likelihood for future use is one of the interesting issues. Mancinelli et al. [21] researched to optimize the selection process of the right donor. They evaluated the effects of factors such as neonatal weight, gender, Apgar score at minute 5, mode of delivery, type of blood sampling, maternal age, and number of deliveries, umbilical cord length, placental weight, and fecal presence in the infant on the qualitative indices such as blood volume, TNC, and CD34+. Solves et al. [22] wrote a paper on the selection process improvement. In this research, 1300 samples were studied and the effects of various factors such as maternal age, gestational age, number of pregnancies, delivery time, neonatal weight, placental weight, mode of delivery, neonatal gender, and type of blood sampling on the qualitative measures such as blood volume, TNC, HTC, CD34+, CFU, and viability were evaluated. Page et al. [23] examined the effects of maternal and neonatal factors as well as the type of blood sampling on the qualitative indices such as TNC, CD34+, and CFU. Manegold et al. [24] researched to reduce the sample rejection rate due to the low amount of cells in them. Wu et al. [25] analyzed the data of 4613 blood samples from Guangzhou's blood bank. In this study, they used statistical tools such as LR, chi-square test, and t-test to analyze the data

Although the previous studies mainly investigated the effective factors and tried to improve the cells selection for cord blood bank, there are no studies in which investigated the quality of the cells considering the effecting factors. In this paper, we predict the cell quality based on the effecting factors using Artificial Neural Networks (ANN) to select the best sample to store in the cord blood bank.

## 3 | Methodology

The classification prediction techniques are one of the most common methods in model learning. The classification is used to find a model to determine the class of objects according to their characteristics. In the classifier algorithms, the initial data set is divided into two sets of training data, and test data. The model is constructed, using the training data set, and the test data are used for validation. In this research, a two-stage method is developed to construct a hybrid intelligent model for the classification prediction of the blood samples.

The first stage is the pre-processing of data. In this stage, two transformation techniques are used. First, nominal and ordinal data are converted to continuous data. Then, all values in each attribute are mapped into the standard interval of [0, 1]. In the second stage, four methods, including MLPNNs, RBFNNs, LR, and C4.5 DT are used for prediction samples quality.

### 3.1 | Data Transformation

Using transformed data is more useful in most heuristic methods, especially when dealing with forecasting problems [26]. According to the structure of the existing attributes in the data sets, the data of qualitative attributes were converted to the data of new quantitative attributes [27]. In this case, a qualitative characteristic will be extended into several quantitative features in which the total of quantitative values will be an index of the qualitative feature. For example, the characteristic of being a man or woman is represented by (0, 1) or (1, 0).

### 3.2 | Data Normalization

Data normalization is used in different forecasting studies, for example, [28], [30]. There are different normalization algorithms, such as Min–Max normalization, Z-score normalization, and sigmoid normalization. In this paper, we use Min–Max normalization. The Min–Max normalization scales the numbers in a data set to improve the accuracy of the subsequent numeric computations. If $X_{old}$, $X_{Max}$ and $X_{Min}$ are the original, maximum and minimum values of the raw data respectively, and $X^*_{Max}, X^*_{Min}$

are the maximum and minimum of the normalized data. New normalized values can be obtained by the following transformation function:

$$X^{*}_{New} = \left( \frac{X_{old} - X_{Min}}{X_{Max} - X_{Min}} \right) (X^{*}_{Max} - X^{*}_{Min}) + X^{*}_{Min}. \quad (1)$$

## 3.3 | Artificial Neural Networks

ANNs are flexible computing frameworks for modeling linear and nonlinear problems [31]. One of the significant advantages of the neural network models is that they can be applied to different classification predictions with high accuracy. This advantage is the result of the power of parallel data processing. Also, no previous assumptions are needed to build the model. ANNs consist of an interconnection of some neurons. There are many varieties of connections under study and here, we discuss two types of network, which are called multilayer perceptron and Radial Basis Function (RBF).

## 3.4 | Multilayer Perceptron

We use a typical three-layer and four-layer feed forward model for the MLP method for classification. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes. The symbolic structure of a MLPNN is shown in *Fig. 1*.

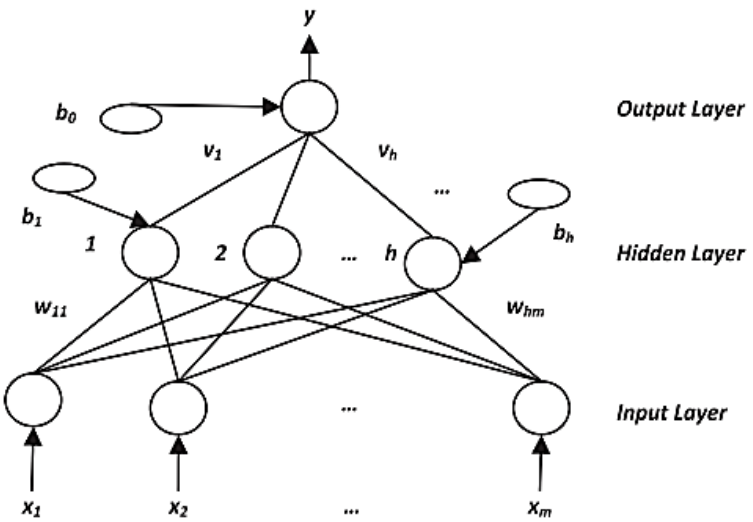


**Fig. 1. MLP neural network model.**

The training and learning process of this network is carried out through this algorithm:

**Step 1.** Initialize weights and thresholds to small random values.

**Step 2.** Choose an input-output pattern $(x^{(k)}, t^{(k)})$ from the training data.

**Step 3.** Compute the network's actual output $(o^{(k)} = f(\sum_{i=1}^{l} w_i x_i^{(k)} - \theta))$, (L is the size of input vector or the size of input neurons.)

Adjust the weight and bias according to the Levenberg-Marquart algorithm.

**Step 4.** If the whole epoch is complete, pass to the following step; otherwise, go to Step 2.

**Step 5.** If the weights and bias reach a steady state $\Delta w_i \approx 0$ through the whole epoch, stop the learning; else go through one more epoch.

The Levenberg Marquardt (LM) algorithm is the most widely used optimization algorithm. LM is similar to error back propagation in which it requires the calculation of the gradient vector, but in addition, LM also computes the Jacobian [32]. The gradient vector is represented as:

$$g = \begin{pmatrix} \frac{\partial E}{\partial W_1} \\ \frac{\partial E}{\partial W_2} \\ \vdots \\ \frac{\partial E}{\partial W_n} \end{pmatrix}, \tag{2}$$

where E is the error of the network for the pattern and W refers to the weights. The Jacobian is calculated as below:

$$J = \begin{bmatrix} \frac{\partial E_1}{\partial W_1} & \frac{\partial E_1}{\partial W_2} & \cdots & \frac{\partial E_1}{\partial W_n} \\ \frac{\partial E_2}{\partial W_1} & \frac{\partial E_2}{\partial W_2} & \cdots & \frac{\partial E_2}{\partial W_n} \\ \frac{\partial E_p}{\partial W_1} & \frac{\partial E_p}{\partial W_2} & \cdots & \frac{\partial E_p}{\partial W_n} \end{bmatrix}. \tag{3}$$

Once the Jacobian is calculated, the LM algorithm can be represented by the following:

$$W_{k+1} = W_k - (J_k^T J_k + \mu I)^{-1} J_k^T E. \tag{4}$$

Where E is the total error for all patterns, I is the identity matrix, and $\mu$ is a learning parameter. The learning parameter $\mu$ is then adjusted several times in each iteration and the result with the greatest reduction of error is selected. When the $\mu$ value is very large, the LM algorithm becomes steepest descent or BP, and when $\mu$ is equal to zero it is the Newton method. The entire process is then repeated until the error is reduced to the required value.

## 3.5 | Radial Basis Function

RBF uses a series of basic functions that are symmetric and centered at each sampling point. *Fig. 2* shows the structure of the RBF. The input neurons have no weight, thus the first hidden layer receives the same values as the first layer. The designed function in the hidden layer is the Radial Basis type. The transfer function for the neurons of the hidden layer is non-monotonic; then the output of these neurons is sent to the output layer by weights. The neurons of the output layers are actually, simple summations.

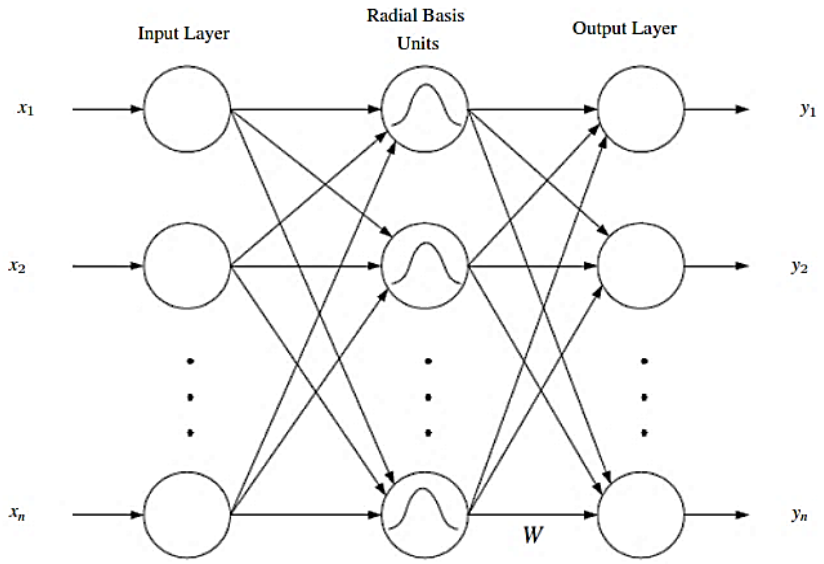


**Fig. 2. Schematic diagram of RBF architecture.**

Let us assume that there are H neurons in the hidden layer. The transfer function is like gaussian density functions. The gaussian function is introduced by the following equation:

$$a_{h,k} = \exp(-\frac{||\hat{x}_h - x_k||^2}{\sigma_h^2}), \tag{5}$$

where $a_{h,k}$ is the output of the h[th] neuron in the hidden layer. Also, $x_h$ is the center of the radial function, and $\sigma$ is the distance scaling parameter.

Finally, the weighted average of the outputs associated with the hidden layer determines the output. In other words, *Eq. (6)* shows the output value.

$$y_i = \sum_{i=1}^{n} w_i \times a_{h,i}, \tag{6}$$

where the $w_i$, is the weight assigned to the i[th] neuron in the hidden layer. Since this method is an observer learning method, the exact values for $x_i$ and $y_i$ are predetermined, thus to have the weights in the second layer, the pseudo-inverse method is used as bellow:

$$G = [g_{i,j}], \tag{7}$$

where

$$g_{i,j} = \exp\left(\frac{-\left\|x_i - v_j\right\|}{2\sigma_j^2}\right) \quad i = 1,2,\ldots,n; j = 1,2,\ldots,p, \tag{8}$$

And we have

$$D = GW, \tag{9}$$

where D is the desired output for the trained data. If $G^{-1}$ exists, then we have

$$W = G^{-1}D. \tag{10}$$

If G is ill-conditioned (close to singularity) or is a non-square matrix, then:

$$W = G^{+}D, \tag{11}$$

where

$$G^{+} = (G^{T}G)^{-1} \times G^{T}. \quad (12)$$

## 3.6 | C4.5

C4.5 is a well-known algorithm used to generate a DT. The C4.5 algorithm improves in DT learning (ID3) regard to the splitting rule and the calculation method [33]. The DTs generated by the C4.5 algorithm can be used for classification. Learned trees can also be represented as sets of if-then rules to improve human readability. C4.5 DT learning is a heuristic, one-step look ahead (hill climbing), non-backtracking search through the space of all possible DTs [34]. The algorithm of C4.5 is shown in the following steps. Training dataset and attributes are introduced as T, and S, respectively.

**Algorithm 1. C4.5 algorithm.**

If T is Null, then
  Return failure
End if
If S is Null, then
  Return Tree as a single node with most frequent class label in T
End if
If |S|=1 then
  Return Tree as a single node
End if
Tree={}
Set
For $a \in S$ do
  Set Info (a,T)=0, and SplitInfo (a,T)=0
  Compute Entropy (a)
    For v ∈ values (a,T) do
      Set $T_{a,v}$ as the subset of T with attribute a=v
Info (a,T) = Info (a,T)+$\frac{|T_{a,v}|}{T_a}$ Entropy ($a_v$)

Spil Info (a, T) = Split Info (a,T)-$\frac{|T_{a,v}|}{T_a}$ log $\frac{|T_{a,v}|}{T_a}$

Gain (a,T) = Entropy (a)-Info (a, T)

Gain Ratio (a,T)=$\frac{\text{Gain (a,T)}}{\text{SplitInfo (a,T)}}$

End for

Set $a_{best}$ = argmax {Gain Ratio (a, T)}

Attach $a_{best}$ into Tree
For v ∈ values ($a_{best}$, T) do
  Call C4.5 ($T_{a,v}$)
End for
Return Tree.

Suppose C denotes the number of classes, and P(S,j) is the proportion of instances in S that are assigned to $j^{th}$ class. Hence, the entropy of attribute S is calculated as follows:

$$\text{Entropy}(S) = -\sum_{j=1}^{C} p(S,j) \times \log p(S,j). \quad (13)$$

Information gain by a training dataset $T$ is defined as:

$$\text{Gain}(S,T) = \text{Entropy}(S) - \sum_{v \in \text{Values}(T_S)} \frac{|T_{S,v}|}{|T_S|} \text{Entropy}(S_v), \quad (14)$$

where Values ($T_s$) is the set of values for S in T, $T_s$ is the subset of T induced by S and $T_{s,v}$ is the subset of T in which attributes S has a value of v.

Therefore, the information gain ratio of attributes S is defined as:

$$\text{Gain Ratio}(S,T) = \frac{\text{Gain}(S,T)}{\text{SplitInfo}(S,T)}, \tag{15}$$

where SplitInfo (S, T) is calculated as:

$$\text{Split Info}(S,T) = -\sum_{v \in \text{Values}(T_S)} \frac{|T_{S,v}|}{|T_S|} \times \log \frac{|T_{S,v}|}{|T_S|}. \tag{16}$$

## 3.7 | Logistic Regression

LR is used as a statistical algorithm for prediction and diagnosis in many disciplines. This model is very effective to solve relatively less complex problems [35]. LR is a regression method for predicting a dichotomous dependent variable. In producing the LR equation, the maximum-likelihood ratio was used to determine the variables' statistical significance [36]. In LR models, the dependent variable is always in categorical form and has two or more levels. Independent variables may be in the numerical or categorical form [37]. We consider the situation where we observe a binary outcome variable y and a vector $x = (1, x_1, x_2, \ldots, x_k)$ of covariates for each N individuals. We code the two-class via a 0/1 response $y_i$, where $y_i=1$ for the first class and $y_i=0$ for the second one. Let P be the conditional probability associated with the first class. LR is a widely used statistical modeling technique in which the probability P of the dichotomous outcome event is related to a set of explanatory variables X in the bellow form:

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = f(X,\beta) = \beta^T X, \tag{17}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ is the vector of the coefficients and $\beta^T$ is the transpose vector. We refer to $p/(1-p)$ as odds-ratio and to the *Eq. (17)* as the log-odds or logit transformation.

Let $D = \{(xi, yi) \mid i = 1, 2, \ldots, n\}$ be the training data set, where the number of samples is n. Here, we assume that the training sample is a realization of a set of independent and identically distributed random variables. The unknown regression coefficients $\beta_i$, which should be estimated from the data, are directly interpretable as log-odds ratios or, in terms of exp($\beta_i$), as odds ratios. That log-likelihood for n observations is:

$$l(\beta) = \sum_{i=1}^{n} \left[ y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right]. \tag{18}$$

The log-likelihood function is used to estimate the regression coefficients $\beta_i$. The exponential value of regression coefficients ($e^{\beta^T}$) gives odds ratio, and this value reflects the effect of risk factor in the disease, and the interpreted values are odds ratios.

# 4 | Evaluation of the Classification Prediction Model

To select the appropriate model, we use three criteria of accuracy, sensitivity, and specificity. The accuracy of a classification prediction model on a given set is the percentage of test records that are correctly identified by the classifier. The accuracy can be calculated using the below formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{19}$$

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

Sensitivity is also considered as the actual positive rate, i.e., the proportion of positive records that are correctly identified. While specificity is the actual negative rate, that is, the ratio of negative records that are correctly identified. In other words:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{20}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{21}$$

## 4.1 | Cross Validation

In order to assess the accuracy of the classification models, a 5-fold cross-validation method has been utilized. In this method, the model is trained and tested five times. First, the data are divided into five sets. In the first step, the first four parts are used for training, and the fifth part is reserved for a test. Then for the second step, the data of parts one to three, and the last part are used for training and the fourth part is for the test. This process is repeated until the stage where the data of parts two to five are used for training, and the first part is used for the test. Finally, the obtained average value indicates the accuracy of the model.

## 4.2 | Data Compilation

The process of data collection was carried out by reviewing the literature. The collection of the important variables that are effective on the output can accelerate the model design, and improve the results. A summarized list of the input and output variables is shown in *Table 2*.

**Table 2. Summary of the literature in the field of the umbilical cord blood sample quality.**

| Article | The Studied Factors | Qualitative Attributes | Sample Size |
|---|---|---|---|
| Solves et al. [22] | Maternal age, number of gravidities, gestational age, neonatal gender, neonatal weight, placental weight, delivery type, delivery duration | TNC | 1300 |
| Nakagawa et al. [38] | Neonatal weight, umbilical cord length, weight of placenta, neonatal gender, gestational age. | (TNC), CD34+ cell, | 956 |
| Jan et al. [11] | Maternity age, size of placenta, fetal weight, number of gravidities, neonatal gender | TNCs, CD34+, CD45+, NRBCs, and viability. | 206 |
| Urciuoli et al. [40] | Gestational age, neonatal weight, placental weight, gender, head diameter, head circumference, abdominal circumference, umbilical cord length, delivery type, blood white cells count | Blood volume, TNC, CD34+, Total CFU count, BFU-E count, CFU-GM count, CFU-GEMM count | 365 |
| Coldwll et al. [41] | Maternal age, neonatal gender, neonatal weight, placental weight, delivery type, meconium n the Amniotic fluid, gestational age, Apgar score, umbilical cord twist around the neck, umbilical cord length, multiparty delivery, maternal diabetes | Volume, TNC, CD34+, HPC recovery | 49 |
| Abdu Wahid et al. [10] | Gestational age, maternal age, neonatal gender, neonatal weight, placental weight, delivery type, Systolic blood pressure | Volume, TNC, CD34+, UCB pH | 47 |
| Wen et al. [19] | Maternal age, neonatal weight, placental weight, delivery type, umbilical cord length, number of gravidities, neonatal gender | Volume, TNC, CD34+ | 1549 |
| Lee et al. [9] | Gestational age, maternal age, neonatal gender, neonatal weight, placental weight, delivery type, neonatal blood type | Volume, TNC, CD34+, Pre viability, Post viability | 11098 |

Table 2. Continued.

| Article | The Studied Factors | Qualitative Attributes | Sample Size |
|---|---|---|---|
| Al-Sweedan et al. [8] | Gestational age, neonatal weight, maternal weight, neonatal platelet cells count, delivery type, incidence of pre-eclampsia, maternal blood pressure, umbilical cord twist around the neck | TNC, CD34+ | 200 |
| Cobellis et al. [20] | Gestational age, neonatal weight, placental weight, gender, head diameter, head circumference, abdominal circumference, umbilical cord length, delivery type | Blood volume, TNC, CD34+ | 219 |
| Keersmaekers et al. [42] | Gestational age, neonatal ethnicity, neonatal gender, neonatal weight, | TNC | 7839 |
| Manegold-Brauer et al. [16] | Maternal age, number of gravidities, maternal height, maternal weight at the beginning and end of gestation, gestational age, neonatal gender, neonatal weight | TNC | 758 |
| Abdelrazik et al. [43] | Weight, gestational age, neonatal gender, delivery type, maternal weight | TNC, CD34+ | 200 |

To summarize, it can be said that the factors related to maternal conditions, neonatal, and delivery conditions are the ones that are considered to be effective on the quality of the umbilical cord blood samples in the past literature. In *Table 3*, some of the sample parameters for each category are given.

Table 3. Categorization of the effective factors, according to the past literature.

| Maternal Factors | Neonatal Factors | Factors of Delivery Conditions |
|---|---|---|
| Maternal age | Birth weight | Delivery type (cesarean section / normal delivery) |
| Number of gravidities | Neonatal gender | Gestational age (weeks of pregnancy) |
| Maternal height | Placental weight | Intrauterine/ extrauterine (extravaginally) blood collection |
| Maternal weight at the beginning and end of pregnancy | Umbilical cord length | |
| Blood pressure | Head diameter | |
| Maternal white blood cells count | Head circumference | |
| Blood type | Thigh bone length | |

The data used in this study were collected from the Electronic Health Record of Royan's umbilical cord blood bank. For the evaluation and quality prediction, the previously used samples in the database are studied. These data were collected from 2012-09-24 to 2013-03-26 and 2015-09-28 to 2016-04-06. According to the available data types, two types of data sets are used. The collected value types are compiled in *Table 4*. The number of data in the first dataset was 71, and the second set contained 618 records. The basic information about these two sets is shown in *Tables 5* and *6*. Both data sets follow the same structure regarding the batch features. *Table 7* shows a list of the batch attributes in each category.

Table 4. The value types used for the data.

| Row | Factors | Type of Values | Type of Variable | Row | Factors | Type of Values | Type of Variable |
|---|---|---|---|---|---|---|---|
| 1 | Maternal age | Discrete | Independent | 12 | Uterine problems | Nominal | Independent |
| 2 | Number of gravidities | Discrete | Independent | 13 | Type of delivery | Nominal | Independent |
| 3 | Duration of pregnancy | Discrete | Independent | 14 | Mother's active disease | Nominal | Independent |
| 4 | Delivery type | Nominal | Independent | 15 | Blood collecting method | Nominal | Independent |
| 5 | Placental exit state | Nominal | Independent | 16 | Punch | Ordinal | Independent |
| 6 | Placental clamp state | Nominal | Independent | 17 | Apgar in 1 minute | Discrete | Independent |
| 7 | Neonatal gender | Nominal | Independent | 18 | Placental weight | Discrete | Independent |
| 8 | Apgar in 5 minutes | Discrete | Independent | 19 | Number of arteries and veins | Nominal | Independent |
| 9 | Connecting position of the umbilical cord | Nominal | Independent | 20 | Umbilical cord length | Discrete | Independent |
| 10 | Birth weight | Discrete | Independent | 21 | Discard or freeze | Nominal | Dependent |
| 11 | Number of pregnancies | Discrete | Independent | | | | |

Table 5. Basic information about set 1.

| Factors | Average | Variance | Standard Deviation | Maximum | Minimum |
|---|---|---|---|---|---|
| Maternal age | 30.88 | 14.57 | 3.81 | 41 | 22 |
| Umbilical cord length | 43.55 | 198.28 | 14.08 | 77 | 15 |
| Number of gravidity | 1.07 | 0.067 | 0.25 | 2 | 1 |
| Duration of pregnancy (weeks) | 39.12 | 116.45 | 10.79 | 41 | 35 |
| Apgar in 5 minutes | 9.85 | 0.12 | 0.35 | 10 | 9 |
| Birth weight | 3212.39 | 167885.3 | 409.73 | 4600 | 2000 |
| Number of pregnancies | 1.44 | 0.45 | 0.673 | 3 | 1 |
| Apgar in 1 minute | 9.05 | 0.16 | 0.41 | 10 | 8 |
| Placental weight | 527.77 | 13132.67 | 114.59 | 839 | 240 |

Table 6. Basic information about set 2.

| Factors | Average | Variance | Standard Deviation | Maximum | Minimum |
|---|---|---|---|---|---|
| Maternal age | 31.14 | 16.91 | 4.11 | 43 | 19 |
| Umbilical cord length | 46.07 | 183.91 | 13.56 | 90 | 10 |
| Number of gravidity | 1.1 | 0.11 | 0.33 | 3 | 1 |
| Duration of pregnancy (weeks) | 39.13 | 1.46 | 1.21 | 42 | 28 |
| Apgar in 5 minutes | 9.85 | 0.14 | 0.37 | 10 | 7 |
| Birth weight | 3141.49 | 120559.6 | 347.21 | 4250 | 2000 |
| Number of pregnancies | 1.56 | 0.69 | 0.83 | 6 | 1 |
| Apgar in 1 minute | 9 | 0.27 | 0.52 | 10 | 3 |
| Placental weight | 519.08 | 17431.62 | 132.02 | 1000 | 250 |

Table 7. Basic information about the categorical features of the data.

| Factors | Number of Categories | Description |
|---|---|---|
| Uterine problems | 2 | Has / Doesn't have |
| Type of delivery | 2 | Normal /Cesarean section |
| Condition of delivery | 2 | Emergency / Expected |
| Mother's active disease | 2 | Healthy / Ill |
| Placental exit state | 3 | With elongation /No intervention /Twisted placenta |
| Blood collection method | 3 | Intrauterine /Ectopic / Both |
| Placental clamp state | 2 | Close to fetus /Close to placenta |
| Punch count | 3 | Once /Twice /More than twice |
| Gender | 2 | Female / Male |
| Connecting position of the umbilical cord | 4 | Curtain /Marginal /Central /Out of center |
| Number of arteries and veins | 2 | One and two / Two and one |
| Discard or freeze | 2 | Discard/ Freeze |

## 5 | Result and Discussion

In this section, the results of the proposed classification methods are presented. To assess the appropriate prediction method, the three criteria of specificity, sensitivity, and accuracy have been considered. All results are expressed separately for sets 1 and 2.

**Set No. 1:** The methods used for this database are discussed in the methodology section. The results of LR and DTs are shown in *Table 8*.

Table 8. The results of the DT and LR.

| Method | Specificity | Sensitivity | Accuracy |
|---|---|---|---|
| LR | 26.66% | 96.66% | 86% |
| DT | 16.6% | 81.2% | 71.46% |

Also the DT extracted for fold 2 is shown in *Fig. 3*.

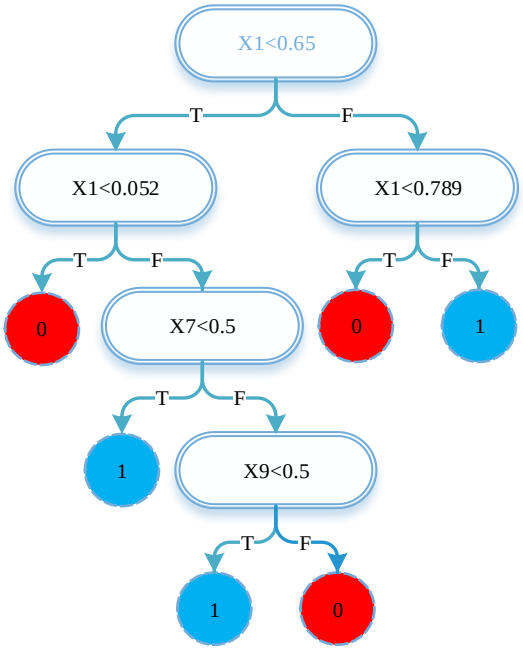


**Fig. 3. The tree schema of generated rules for fold 2 prediction of database 1.**

Different designs were proposed to use the RBFNNs. In order to adjust the input parameters such as the center and radius, the k-means clustering method is issued. 16 different designs are considered for the RBF neural network. *Table 9* shows the results of RBFNNs. It can be seen that the highest accuracy in these methods was achieved using 40 clusters, which is about 91.5%. In terms of the sensitivity criteria, several different designs have managed to reach as high as 95 percent. In design No. 10, using 20 neurons in the hidden layer, the highest value of specificity was achieved compared with other designs. Ultimately, design No. 14 has the best performance with 40 neurons in the hidden layer.

**Table 9. The results of the radial basis neural network method.**

| Design NO. | Number of Clusters (Neurons) | Activation Function | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|
| 1 | 3 | Gaussian | 60% | 85% | 81.9% |
| 2 | 4 | Gaussian | 50% | 91.7% | 86.1% |
| 3 | 5 | Gaussian | 56.7% | 95% | 88.9% |
| 4 | 6 | Gaussian | 80% | 90% | 88.9% |
| 5 | 7 | Gaussian | 63.3% | 91.7% | 87.4% |
| 6 | 8 | Gaussian | 53.3% | 93.3% | 87.4% |
| 7 | 9 | Gaussian | 66.7% | 91.7% | 87.5% |
| 8 | 10 | Gaussian | 60% | 95% | 90.1% |
| 9 | 15 | Gaussian | 53.3% | 95% | 88.8% |
| 10 | 20 | Gaussian | 90% | 86.7% | 87.3% |
| 11 | 25 | Gaussian | 73.3% | 90% | 87.3% |
| 12 | 30 | Gaussian | 63.3% | 95% | 90.2% |
| 13 | 35 | Gaussian | 73.3% | 93.3% | 90.2% |
| 14 | 40 | Gaussian | 73.3% | 95% | 91.5% |
| 15 | 45 | Gaussian | 63.3% | 95% | 90.2% |
| 16 | 50 | Gaussian | 73.3% | 91.7% | 88.8% |

*Table 10* shows the results of the MLPNNs method. In the construction of these networks, one or two hidden layers with two different activation functions were used. The number of neurons in each hidden layer is shown in *Table 10*. 66 different designs of the MLPNN were constructed and evaluated.

Table 10. The results of the MLPNNs method.

| Design NO. | Number of Hidden Layers | Neurons in each Hidden Layer | Activation Function | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | Sigmoid | 26.6% | 76.2% | 69.6% |
| 2 | 1 | 3 | Sigmoid | 60% | 76.7% | 73.1% |
| 3 | 1 | 4 | Sigmoid | 30% | 78.3% | 70.4% |
| 4 | 1 | 5 | Sigmoid | 26.6% | 85% | 76.1% |
| 5 | 1 | 6 | Sigmoid | 30% | 80% | 71.8% |
| 6 | 1 | 7 | Sigmoid | 36.7% | 78.3% | 71.7% |
| 7 | 1 | 8 | Sigmoid | 20% | 81.7% | 71.9% |
| 8 | 1 | 9 | Sigmoid | 46.7% | 83.3% | 77.4% |
| 9 | 1 | 2 | Tanh | 40% | 63.3% | 58.9% |
| 10 | 1 | 3 | Tanh | 30% | 61.7% | 56.3% |
| 11 | 1 | 4 | Tanh | 50% | 59.3% | 54.9% |
| 12 | 1 | 5 | Tanh | 46.7% | 51.7% | 51% |
| 13 | 1 | 6 | Tanh | 36.7% | 61.7% | 60% |
| 14 | 1 | 7 | Tanh | 30% | 78.3% | 70.4% |
| 15 | 1 | 8 | Tanh | 16.7% | 75% | 66.2% |
| 16 | 1 | 9 | Tanh | 6.7% | 73.3% | 63.2% |
| 17 | 2 | (1:2,2:2) | Sigmoid | 46.7% | 75% | 70.3% |
| 18 | 2 | (1:2,2:3) | Sigmoid | 20% | 88.4% | 77.8% |
| 19 | 2 | (1:2,2:4) | Sigmoid | 26.7% | 70% | 63.1% |
| 20 | 2 | (1:2,2:5) | Sigmoid | 36.7% | 75% | 69.2% |
| 21 | 2 | (1:2,2:6) | Sigmoid | 36.7% | 67.2% | 59% |
| 22 | 2 | (1:3,2:2) | Sigmoid | 20% | 80% | 70.5% |
| 23 | 2 | (1:3,2:3) | Sigmoid | 20% | 78.3% | 69.1% |
| 24 | 2 | (1:3,2:4) | Sigmoid | 36.7% | 78.3% | 71.6% |
| 25 | 2 | (1:3,2:5) | Sigmoid | 26.7% | 81.7% | 73.2% |
| 26 | 2 | (1:3,2:6) | Sigmoid | 26.7% | 76.7% | 68.8% |
| 27 | 2 | (1:4,2:2) | Sigmoid | 10% | 80% | 68.9% |
| 28 | 2 | (1:4,2:3) | Sigmoid | 26.7% | 80% | 71.8% |
| 29 | 2 | (1:4,2:4) | Sigmoid | 36.7% | 83.3% | 75.9% |
| 30 | 2 | (1:4,2:5) | Sigmoid | 30% | 71.6% | 64.6% |
| 31 | 2 | (1:4,2:6) | Sigmoid | 10% | 83.3% | 71.8% |
| 32 | 2 | (1:5,2:2) | Sigmoid | 30% | 85% | 76% |
| 33 | 2 | (1:5,2:3) | Sigmoid | 26.7% | 85% | 76% |
| 34 | 2 | (1:5,2:4) | Sigmoid | 36.7% | 88.3% | 80.3% |
| 35 | 2 | (1:5,2:5) | Sigmoid | 20% | 81.7% | 71.7% |
| 36 | 2 | (1:5,2:6) | Sigmoid | 26.7% | 83.3% | 74.5% |
| 37 | 2 | (1:6,2:2) | Sigmoid | 40% | 83.3% | 76% |
| 38 | 2 | (1:6,2:3) | Sigmoid | 30% | 81.7% | 73.2% |
| 39 | 2 | (1:6,2:4) | Sigmoid | 30% | 85% | 76% |
| 40 | 2 | (1:6,2:5) | Sigmoid | 26.7% | 75% | 67.1% |
| 41 | 2 | (1:6,2:6) | Sigmoid | 36.7% | 73.3% | 67.5% |
| 42 | 2 | (1:2,2:2) | Tanh | 43.3% | 65% | 62% |
| 43 | 2 | (1:2,2:3) | Tanh | 10% | 73.3% | 63.2% |
| 44 | 2 | (1:2,2:4) | Tanh | 46.7% | 71.7% | 67.8% |
| 45 | 2 | (1:2,2:5) | Tanh | 56.7% | 58.3% | 57.9% |
| 46 | 2 | (1:2,2:6) | Tanh | 26.7% | 75% | 67.7% |
| 47 | 2 | (1:3,2:2) | Tanh | 20% | 83.3% | 73.3% |
| 48 | 2 | (1:3,2:3) | Tanh | 50% | 71.7% | 67.7% |
| 49 | 2 | (1:3,2:4) | Tanh | 33.3% | 73.3% | 67.7% |
| 50 | 2 | (1:3,2:5) | Tanh | 16.7% | 70% | 61.9% |
| 51 | 2 | (1:3,2:6) | Tanh | 56.7% | 70% | 67.8% |
| 52 | 2 | (1:4,2:2) | Tanh | 63.3% | 61.7% | 62% |
| 53 | 2 | (1:4,2:3) | Tanh | 10% | 85% | 73.2% |
| 54 | 2 | (1:4,2:4) | Tanh | 36.7% | 78.3% | 71.6% |
| 55 | 2 | (1:4,2:5) | Tanh | 56.7% | 61.7% | 60.7% |
| 56 | 2 | (1:4,2:6) | Tanh | 36.7% | 63.3% | 59.1% |
| 57 | 2 | (1:5,2:2) | Tanh | 36.7% | 61.7% | 57.8% |
| 58 | 2 | (1:5,2:3) | Tanh | 30% | 71.7% | 64.8% |
| 59 | 2 | (1:5,2:4) | Tanh | 10% | 81.7% | 70.5% |

Table 10. Continued.

| Design NO. | Number of Hidden Layers | Neurons in each Hidden Layer | Activation Function | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|---|
| 60 | 2 | (1:5,2:5) | Tanh | 10% | 75% | 64.7% |
| 61 | 2 | (1:5,2:6) | Tanh | 33.3% | 71.7% | 66% |
| 62 | 2 | (1:6,2:2) | Tanh | 16.7% | 88.3% | 77.4% |
| 63 | 2 | (1:6,2:3) | Tanh | 16.7% | 78.3% | 68.9% |
| 64 | 2 | (1:6,2:4) | Tanh | 33.3% | 73.3% | 67.3% |
| 65 | 2 | (1:6,2:5) | Tanh | 53.3% | 76.6% | 73.3% |
| 66 | 2 | (1:6,2:6) | Tanh | 26.7% | 86.7% | 77.5% |

The best results in specificity, sensitivity, and accuracy were achieved for designs No. 52, 18, and 34, respectively. In order to predict the discarding or freezing of samples, four methods of LR, C4.5 DT, RBFNNs (hybrid with the k-means method), and MLPNNs were used. The maximum value of sensitivity was achieved using the LR method, but regarding the two criteria of specificity and accuracy, RBFNN method has the highest scores.

**Set No. 2:** The results of LR and decision trees are shown in *Table 11*.

Table 11. The results of the DT and LR.

| Data Set | Method | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|
| 2 | LR | 35.27% | 83.94% | 73.62% |
| 2 | DT | 36.2% | 86.8% | 77.6% |

Also, the DT extracted for fold 2 is shown in *Fig. 4*.

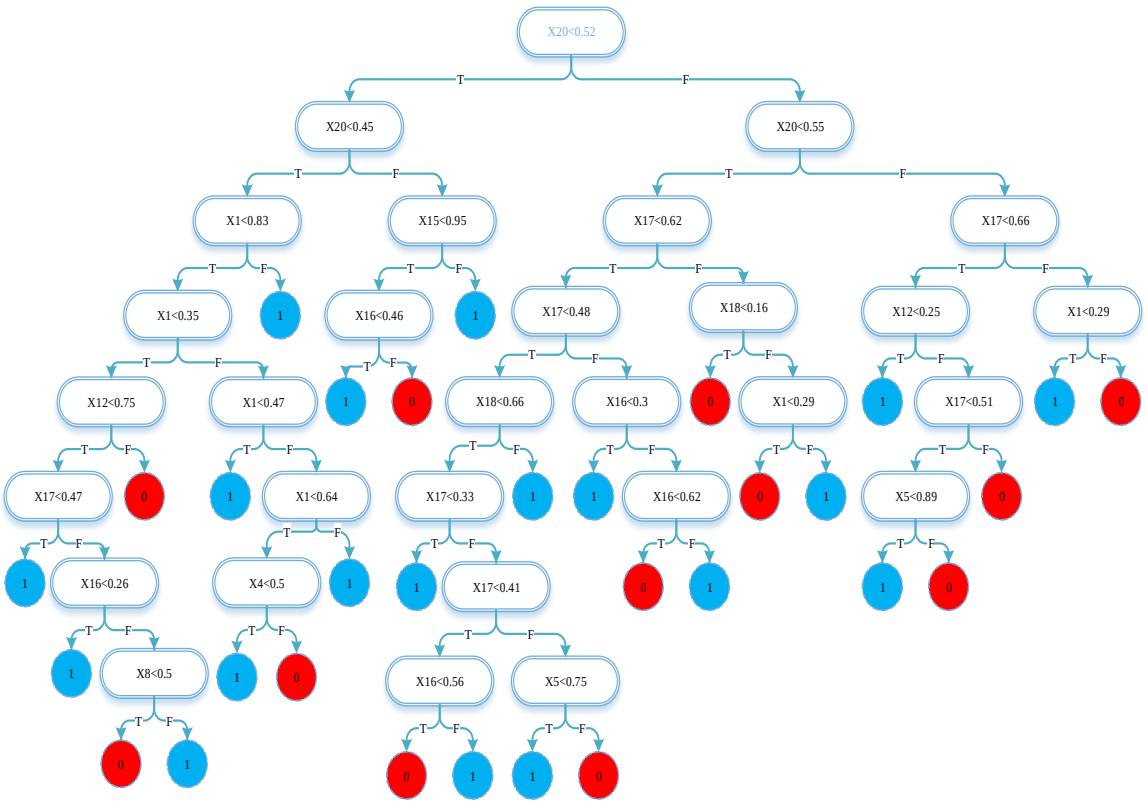


Fig. 4. Tree schema of generated rules for fold 2 prediction, using the DT method.

Different designs were proposed to use the RBFNNs. In order to adjust the input parameters such as the center and radius, the k-means clustering method is issued. 16 different designs are considered for the RBF neural network. *Table 12* shows the results of RBFNNs. It can be seen that the highest accuracy in these methods was achieved in design 3, using 5 clusters which are about 81.6%. In terms of the sensitivity criteria, the 5th design with 7 neurons in the hidden layer has managed to reach as high as 98.8

percent. In design no. 9, using 15 neurons in the hidden layer, the highest value of specificity was achieved compared with other designs.

**Table 12. The results of the radial basis neural network method.**

| Design NO | Number of Clusters | Activation Function | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|
| 1 | 3 | Gaussian | 11.7% | 95.8% | 80.4% |
| 2 | 4 | Gaussian | 9.2% | 95.8% | 79.8% |
| 3 | 5 | Gaussian | 10.5% | 97.8% | 81.6% |
| 4 | 6 | Gaussian | 9.8% | 96.9% | 80.5% |
| 5 | 7 | Gaussian | 3.7% | 98.8% | 81.3% |
| 6 | 8 | Gaussian | 12.1% | 93.7% | 78.2% |
| 7 | 9 | Gaussian | 9.9% | 97.1% | 80.7% |
| 8 | 10 | Gaussian | 8.2% | 96.6% | 80.5% |
| 9 | 15 | Gaussian | 25.1% | 93.6% | 81.3% |
| 10 | 20 | Gaussian | 11% | 96.5% | 80.5% |
| 11 | 25 | Gaussian | 12.2% | 96.2% | 80.5% |
| 12 | 30 | Gaussian | 7.7% | 97.6% | 81% |
| 13 | 35 | Gaussian | 5.9% | 98% | 81% |
| 14 | 40 | Gaussian | 8.1% | 95.8% | 79.7% |
| 15 | 45 | Gaussian | 11% | 96.6% | 80.5% |
| 16 | 50 | Gaussian | 9.8% | 96.4% | 80.3% |

*Table 13* shows the results of the MLPNNs method. In the construction of these networks, one or two hidden layers with two different activation functions were used. The number of neurons in each hidden layer is shown in *Table 13*. 66 different designs of MLPNN were constructed and evaluated. The best results in terms of specificity, sensitivity, and accuracy were achieved for design numbers 65 and 9, respectively.

**Table 13. The results of the MLPNNs method.**

| Design NO. | Number of Hidden Layers | Neurons in each Hidden Layer | Activation Function | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | Sigmoid | 26.6% | 76.2% | 68.4% |
| 2 | 1 | 3 | Sigmoid | 28.6% | 85.6% | 75.1% |
| 3 | 1 | 4 | Sigmoid | 24.3% | 80.7% | 70.3% |
| 4 | 1 | 5 | Sigmoid | 23.6% | 84.7% | 73.3% |
| 5 | 1 | 6 | Sigmoid | 19.2% | 79.3% | 68.6% |
| 6 | 1 | 7 | Sigmoid | 24.5% | 79.3% | 69.3% |
| 7 | 1 | 8 | Sigmoid | 33.2% | 77% | 68.5% |
| 8 | 1 | 9 | Sigmoid | 18.8% | 78.7% | 67.7% |
| 9 | 1 | 2 | Tanh | 4.5% | 99.2% | 81.6% |
| 10 | 1 | 3 | Tanh | 16.4% | 88.9% | 75.6% |
| 11 | 1 | 4 | Tanh | 14.3% | 81.2% | 73.6% |
| 12 | 1 | 5 | Tanh | 18.9% | 80.4% | 69.4% |
| 13 | 1 | 6 | Tanh | 28.1% | 71% | 63.2% |
| 14 | 1 | 7 | Tanh | 35.7% | 79% | 71% |
| 15 | 1 | 8 | Tanh | 32.3% | 73% | 65.1% |
| 16 | 1 | 9 | Tanh | 28.1% | 78.1% | 68.6% |
| 17 | 2 | (1:2,2:2) | Sigmoid | 14.9% | 89.2% | 75.4% |
| 18 | 2 | (1:2,2:3) | Sigmoid | 15.7% | 87.2% | 73.8% |
| 19 | 2 | (1:2,2:4) | Sigmoid | 10.8% | 89.7% | 75.4% |
| 20 | 2 | (1:2,2:5) | Sigmoid | 6.1% | 92.8% | 77% |
| 21 | 2 | (1:2,2:6) | Sigmoid | 22.5% | 86.4% | 74.6% |
| 22 | 2 | (1:3,2:2) | Sigmoid | 13.5% | 91.7% | 76.9% |
| 23 | 2 | (1:3,2:3) | Sigmoid | 23.1% | 84.4% | 73% |
| 24 | 2 | (1:3,2:4) | Sigmoid | 15.5% | 86.7% | 73.6% |
| 25 | 2 | (1:3,2:5) | Sigmoid | 21.5% | 85.4% | 73.6% |
| 26 | 2 | (1:3,2:6) | Sigmoid | 27.5% | 79.9% | 70.7% |
| 27 | 2 | (1:4,2:2) | Sigmoid | 25.5% | 79.5% | 69.7% |
| 28 | 2 | (1:4,2:3) | Sigmoid | 27.9% | 81.7% | 71.6% |
| 29 | 2 | (1:4,2:4) | Sigmoid | 19.6% | 78.4% | 67.9% |
| 30 | 2 | (1:4,2:5) | Sigmoid | 20.3% | 81.2% | 69.9% |
| 31 | 2 | (1:4,2:6) | Sigmoid | 25.4% | 82.9% | 72.3% |

Table 13. Continued.

| Design NO. | Number of Hidden Layers | Neurons in each Hidden Layer | Activation Function | Specificity | Sensitivity | Accuracy |
|---|---|---|---|---|---|---|
| 32 | 2 | (1:5,2:2) | Sigmoid | 30.2% | 80.9% | 71.3% |
| 33 | 2 | (1:5,2:3) | Sigmoid | 29.5% | 78.3% | 68.7% |
| 34 | 2 | (1:5,2:4) | Sigmoid | 30.7% | 77.4% | 68.7% |
| 35 | 2 | (1:5,2:5) | Sigmoid | 33% | 75.7% | 67.8% |
| 36 | 2 | (1:5,2:6) | Sigmoid | 21.6% | 81% | 69.7% |
| 37 | 2 | (1:6,2:2) | Sigmoid | 25.7% | 81.3% | 71.8% |
| 38 | 2 | (1:6,2:3) | Sigmoid | 30.9% | 75.6% | 67.2% |
| 39 | 2 | (1:6,2:4) | Sigmoid | 34.8% | 81% | 72.5% |
| 40 | 2 | (1:6,2:5) | Sigmoid | 30.5% | 76.2% | 67.6% |
| 41 | 2 | (1:6,2:6) | Sigmoid | 27% | 76.3% | 67.2% |
| 42 | 2 | (1:2,2:2) | Tanh | 9.7% | 93.5% | 79% |
| 43 | 2 | (1:2,2:3) | Tanh | 8.9% | 90% | 75% |
| 44 | 2 | (1:2,2:4) | Tanh | 16% | 91.5% | 77.4% |
| 45 | 2 | (1:2,2:5) | Tanh | 14.5% | 88.8% | 75.1% |
| 46 | 2 | (1:2,2:6) | Tanh | 16.2% | 81.5% | 69.7% |
| 47 | 2 | (1:3,2:2) | Tanh | 3.2% | 94.9% | 77.8% |
| 48 | 2 | (1:3,2:3) | Tanh | 21% | 78.2% | 67.6% |
| 49 | 2 | (1:3,2:4) | Tanh | 12.5% | 89% | 75% |
| 50 | 2 | (1:3,2:5) | Tanh | 25.5% | 83.1% | 72.5% |
| 51 | 2 | (1:3,2:6) | Tanh | 19.6% | 81.4% | 70% |
| 52 | 2 | (1:4,2:2) | Tanh | 26.5% | 79.4% | 69.7% |
| 53 | 2 | (1:4,2:3) | Tanh | 22% | 81.3% | 70.3% |
| 54 | 2 | (1:4,2:4) | Tanh | 20.9% | 80.3% | 69.3% |
| 55 | 2 | (1:4,2:5) | Tanh | 30.9% | 76.7% | 68.4% |
| 56 | 2 | (1:4,2:6) | Tanh | 27% | 79.8% | 70% |
| 57 | 2 | (1:5,2:2) | Tanh | 30.5% | 74.5% | 66.3% |
| 58 | 2 | (1:5,2:3) | Tanh | 27.4% | 76% | 66.9% |
| 59 | 2 | (1:5,2:4) | Tanh | 25.5% | 79.2% | 69% |
| 60 | 2 | (1:5,2:5) | Tanh | 21.1% | 78.4% | 67.7% |
| 61 | 2 | (1:5,2:6) | Tanh | 17.9% | 77.7% | 66.9% |
| 62 | 2 | (1:6,2:2) | Tanh | 30.4% | 72.8% | 64.2% |
| 63 | 2 | (1:6,2:3) | Tanh | 34.9% | 77.2% | 69.2% |
| 64 | 2 | (1:6,2:4) | Tanh | 28% | 77.9% | 69% |
| 65 | 2 | (1:6,2:5) | Tanh | 36.4% | 74.5% | 67.5% |
| 66 | 2 | (1:6,2:6) | Tanh | 33% | 75.3% | 67% |

In order to predict the discarding or freezing of samples, four methods of LR, C4.5 DT, RBFNNs (hybrid with the k-means method), and MLPNNs were used. The highest values for the sensitivity and specificity criteria were obtained using the MLPNN method. But regarding the accuracy criterion, both methods of MLPNN and RBFNN (hybrid with the k-means method) earned the highest scores.

According to the presented results in the previous section, based on the prediction model of the RBFNN method, the reduced costs of discard and lost opportunity of the collected samples in Royan's umbilical cord blood bank during the past year can be calculated by the following procedure.

The cost of each contract is 22,000,000 Rials, in case of cancellation, 80% of the contract cost will be returned to the referring person. The cost imposed on the company will be different based on the fact that at what stage of the tests, the discard is determined. Hence the average cost of collecting the samples and related tests is estimated to be about 6,000,000 Rials.

The total number of contracts in the umbilical cord blood bank was 11,750 in 2015.

The percentage of contract discard was about 14% in 2015.

Considering the above data, the total reduction cost for sets 1 and 2 are shown in *Table 14*.

Table 14. The total reduction in the contract discard costs (Rials).

| Set 1 | Set 2 |
|---|---|
| $11750*0.14*0.915*(0.8*22000000+6000000)$ $=35522130000$ | $11750*0.14*0.816*(0.8*22000000+6000000)$ $=31678752000$ |

This cost reduction for the company is estimated without considering the lost opportunity costs for discarded contracts, if it's taken into account, the amount of total cost reduction will be more than the above value.

## 6 | Conclusion

Choosing the best umbilical cord blood stem cells for storage, is one of the important issues that needs a proper pre-determined method. In the absence of this method, high costs are imposed on companies, and the optimal use possibility of the stored cells is reduced. In this paper, we proposed a proper classification prediction method (hybrid of radial basis neural network with k-means clustering) for freezing or discarding of umbilical cord blood stem cells, since some frozen records are useless when we need to use them for transplantation. Using the proposed method reduced storage costs and increased the likelihood of cells effectiveness. We implemented the proposed model in Royan institute and saved 1 million dollars for the first year and the provided result showed the method's effectiveness.

## References

[1] Lin, H. (2008). Cell biology of stem cells: an enigma of asymmetry and self-renewal. *The journal of cell biology*, *180*(2), 257-260.

[2] Molofsky, A. V., Pardal, R., & Morrison, S. J. (2004). Diverse mechanisms regulate stem cell self-renewal. *Current opinion in cell biology*, *16*(6), 700-707.

[3] Samuel, G. N., Kerridge, I. H., & O'Brien, T. A. (2008). Umbilical cord blood banking: public good or private benefit? *Medical journal of Australia*, *188*(9), 533-535.

[4] Wadlow, R. C., & Porter, D. L. (2002). Umbilical cord blood transplantation: where do we stand? *Biology of blood and marrow transplantation*, *8*(12), 637-647.

[5] Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell*, *173*(7), 1581-1592.

[6] Jerby-Arnon, L., Pfetzer, N., Waldman, Y. Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., & Clemons, P. A. (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, *158*(5), 1199-1209.

[7] Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., Kamińska, B., Huelsken, J., Omberg, L., & Gevaert, O. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*, *173*(2), 338-354.

[8] Al-Sweedan, S. A., Musalam, L., & Obeidat, B. (2013). Factors predicting the hematopoietic stem cells content of the umbilical cord blood. *Transfusion and apheresis science*, *48*(2), 247-252.

[9] Lee, H. R., Park, J. S., Shin, S., Roh, E. Y., Yoon, J. H., Han, K. S., Kim, B. J., Storms, R. W., & Chao, N. J. (2012). Increased numbers of total nucleated and CD34+ cells in blood group O cord blood: an analysis of neonatal innate factors in the Korean population. *Transfusion*, *52*(1), 76-81.

[10] Abdul Wahid, F. S., Nasaruddin, M. Z., Mohd Idris, M. R., Tusimin, M., Tumian, N. R., & Abdullah Mahdy, Z. (2012). Effects of preeclampsia on the yield of hematopoietic stem cells obtained from umbilical cord blood at delivery. *Journal of obstetrics and gynaecology research*, *38*(3), 490-497.

[11] Jan, R. H., Wen, S. H., Shyr, M. H., & Chiang, B. L. (2008). Impact of maternal and neonatal factors on CD34+ cell count, total nucleated cells, and volume of cord blood. *Pediatric transplantation*, *12*(8), 868-873.

[12] Ballen, K. K., Wilson, M., Wuu, J., Ceredona, A., Hsieh, C., Stewart, F., Popovsky, M., & Quesenberry, P. J. (2001). Bigger is better: maternal and neonatal predictors of hematopoietic potential of umbilical cord blood units. *Bone marrow transplantation*, *27*(1), 7-14.

[13] Surbek, D. V., Visca, E., Steinmann, C., Tichelli, A., Schatt, S., Hahn, S., Gratwohl, A., & Holzgreve, W. (2000). Umbilical cord blood collection before placental delivery during cesarean delivery increases cord blood volume and nucleated cell number available for transplantation. *American journal of obstetrics & gynecology*, *183*(1), 218-221.

[14] Al-Deghaither, S. Y. (2015). Impact of maternal and neonatal factors on parameters of hematopoietic potential in umbilical cord blood. *Saudi medical journal*, *36*(6), 704-712.

[15] Nunes, R. D. & Zandavalli, F. M. (2015). Association between maternal and fetal factors and quality of cord blood as a source of stem cells. *Revista brasileira de hematologia e hemoterapia*, *37*(1), 38-42.
[16] Manegold-Brauer, G., Borner, B., Bucher, C., Hoesli, I., Passweg, J., Girsberger, S., Schoetzau, A., Gisin, S., & Visca, E. (2014). A prenatal prediction model for total nucleated cell count increases the efficacy of umbilical cord blood banking. *Transfusion*, *54*(11), 2946-2952.
[17] Donaldson, C., Armitage, W. J., Laundy, V., Barron, C., Buchanan, R., Webster, J., Bradley, B., & Hows, J. (1999). Impact of obstetric factors on cord blood donation for transplantation. *British journal of haematology*, *106*(1), 128-132.
[18] Jaime-Pérez, J. C., Monreal-Robles, R., Rodríguez-Romo, L. N., Mancías-Guerra, C., Herrera-Garza, J. L., & Gómez-Almaguer, D. (2011). Evaluation of volume and total nucleated cell count as cord blood selection parameters: a receiver operating characteristic curve modeling approach. *American journal of clinical pathology*, *136*(5), 721-726.
[19] Wen, S.-H., Zhao, W.-L., Lin, P.-Y., & Yang, K.-L. (2012). Associations among birth weight, placental weight, gestational period and product quality indicators of umbilical cord blood units. *Transfusion and apheresis science*, *46*(1), 39-45.
[20] Cobellis, L., Castaldi, M., Trabucco, E., Imparato, G., Perricone, F., Frega, V., Signoriello, G., & Colacurci, N. (2013). Cord blood unit bankability can be predicted by prenatal sonographic parameters. *European journal of obstetrics and gynecology and reproductive biology*, *170*(2), 391-395.
[21] Mancinelli, F., Tamburini, A., Spagnoli, A., Malerba, C., Suppo, G., Lasorella, R., De Fabritiis, P., & Calugi, A. (2006). Optimizing umbilical cord blood collection: impact of obstetric factors versus quality of cord blood units. *Transplantation proceedings*, *38*(4), 1174-1176.
[22] Solves, P., Perales, A., Mirabet, V., Blasco, I., Blanquer, A., Planelles, D., Larrea, L., Monleon, J., Carbonell-Uberos, F., & Angeles Soler, M. (2004). Optimizing donor selection in a cord blood bank. *European journal of haematology*, *72*(2), 107-112.
[23] Page, K. M., Mendizabal, A., Betz-Stablein, B., Wease, S., Shoulars, K., Gentry, T., Prasad, V. K., Sun, J., Carter, S., & Balber, A. E. (2014). Optimizing donor selection for public cord blood banking: influence of maternal, infant, & collection characteristics on cord blood unit quality. *Transfusion*, *54*(2), 340-352.
[24] Manegold, G., Meyer-Monard, S., Tichelli, A., Pauli, D., Holzgreve, W., & Troeger, C. (2008). Cesarean section due to fetal distress increases the number of stem cells in umbilical cord blood. *Transfusion*, *48*(5), 871-876.
[25] Wu, S., Xie, G., Wu, J., Chen, J., Lu, Y., Li, Y., Tang, X., & Liao, C. (2015). Influence of maternal, infant, & collection characteristics on high-quality cord blood units in Guangzhou Cord Blood Bank. *Transfusion*, *55*(9), 2158-2167.
[26] Azadeh, A., Saberi, M., Ghaderi, S., Gitiforouz, A., & Ebrahimipour, V. (2008). Improved estimation of electricity demand function by integration of fuzzy system and data mining approach. *Energy conversion and management*, *49*(8), 2165-2177.
[27] Ashoori, M., & Tarokh, M. J. (2014). Applying the process mining project methodology for insurance risks reduction. *International journal of research in industrial engineering*, *3*(4), 57-69.
[28] Azadeh, A., Asadzadeh, S., & Ghanbari, A. (2010). An adaptive network-based fuzzy inference system for short-term natural gas demand estimation: uncertain and complex environments. *Energy policy*, *38*(3), 1529-1536.
[29] Oliveira, A. L. & Meira, S. R. (2006). Detecting novelties in time series through neural networks forecasting with robust confidence intervals. *Neurocomputing*, *70*(1-3), 79-92.
[30] Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., & Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series. *Engineering applications of artificial intelligence*, *17*(2), 159-167.
[31] Jaberi, M., Najafi, S. E., Hoseinzadeh Lotfi, F., & Haji molana, M. (2021). A comprehensive hybrid Ndea-Bsc model and a new neural network for predicting organizational performance indicators. *Journal of decisions and operations research*, *6*(2), 271-287. (**In Persian**). https://doi.org/10.22105/dmor.2020.254632.1243
[32] Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE transactions on neural networks*, *5*(6), 989-993.
[33] Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers http://portal.acm.org/citation.cfm?id=152181
[34] Mitchell, T. M. (2007). *Machine learning* (Vol. 1). New York: McGraw-Hill.

[35] Pal, S., & Gauri, S. (2018). Simultaneous optimization of quantitative and ordinal responses using Taguchi method. *International journal of research in industrial engineering*, *7*(2), 184-205.

[36] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[37] Camdeviren, H. A., Yazici, A. C., Akkus, Z., Bugdayci, R., & Sungur, M. A. (2007). Comparison of logistic regression model and classification tree: an application to postpartum depression data. *Expert systems with applications*, *32*(4), 987-994.

[38] Nakagawa, R., Watanabe, T., Kawano, Y., Kanai, S., Suzuya, H., Kaneko, M., Watanabe, H., Okamoto, Y., Kuroda, Y., & Nakayama, T. (2004). Analysis of maternal and neonatal factors that influence the nucleated and CD34+ cell yield for cord blood banking. *Transfusion*, *44*(2), 262-267.

[39] Wu, J., Liao, C., Chen, J., Xu, Z., Gu, S., Wu, S., Lu, Y., & Xie, G. (2010). Analysis of maternal and neonatal factors associated with hematopoietic reconstruction potential in umbilical cord blood units. *Zhongguo shi yan xue ye xue za zhi*, *18*(6), 1535-1541.

[40] Urciuoli, P., Passeri, S., Ceccarelli, F., Luchetti, B., Paolicchi, A., Lapi, S., Nocchi, F., Lamanna, R., Iorio, M., & Vanacore, R. (2010). Pre-birth selection of umbilical cord blood donors. *Blood transfusion*, *8*(1), 36-43.

[41] Coldwell, K. E., Lee, S. J., Kean, J., Khoo, C. P., Tsaknakis, G., Smythe, J., & Watt, S. M. (2011). Effects of obstetric factors and storage temperatures on the yield of endothelial colony forming cells from umbilical cord blood. *Angiogenesis*, *14*(3), 381-392.

[42] Keersmaekers, C. L., Mason, B. A., Keersmaekers, J., Ponzini, M., & Mlynarek, R. A. (2014). Factors affecting umbilical cord blood stem cell suitability for transplantation in an in utero collection program. *Transfusion*, *54*(3), 545-549.

[43] Abdelrazik, A. M., El Said, M. N., Abdelaziz, H. E. M., Badran, H. M., & Elal, E. Y. A. A. (2015). The impact of fetal and maternal physiologic factors on umbilical cord blood quality as a source of stem cells in Egyptian population. *Transfusion*, *55*(12), 2882-2889.