Statistical Analysis of Hematological Parameters for Prediction of Sickle Cell Disease

Bhawna Dash(✉), Soumyalatha Naveen, and UM Ashwinkumar

School of CSE, REVA University, Bangalore, India
dashbhawna2000@gmail.com

Abstract. About 30 million people worldwide are affected by the monogenic recessive -globin gene abnormality known as sickle cell disease (SCD), which is a significant public health issue. From asymptomatic to severely symptomatic illnesses that might cause patient mortality, pathological features range. The most common presenting symptom of SCD is vasooclussive crisis (VOC). The red cell membrane of the Sickle Red Blood Cells (SRBCs) is damaged by repeated cycles of sickling and desickling processes caused by the formation and aggregation of HbS (sickle hemoglobin) polymers. Cellular dehydration (reduction of ion and water content), increased viscosity (red cell density) and a transient increase in intracellular calcium are all associated with HbS polymerization. As a result, SRBCs become adhesive and inflexible (rigid), resulting in premature destruction. The decreased life span of SRBCs causes chronic hemolytic anemia, and capillary blockage causes tissue hypoxia and subsequent organ damage. So, it is important to monitor patients suffering from sickle cells.

Here we have used machine learning to visualize those patients and categorize them according to their hemoglobin level, percentage of reticulocyte count and serum Lactate dehydrogenase (LDH) level which is regarded as a marker of hemolysis. In this article we propose a framework which uses the statistical analysis using Linear Regression technique on a sickle cell patients dataset showing how hemoglobin is depleted in a body by the use of two parameters called LDH and Retics.

Keywords: Sickle Cell Disease · RBC · WBC · Hemoglobin · Reticulocyte · Machine learning

1 Introduction

Hemoglobinopathy, one of the most prevalent monogenic disorders affecting humans, is responsible for some of the serious genetic and social health difficulties in India, South Africa, Saudi Arabia, South America, and other South Asian and African nations (see Fig. 1). In the history of haemoglobinopathies, sickle cell disease (SCD) is one of the oldest recognized molecular disorder [HbS ($HBB^{Glu6Val}$)] whereas the HbE disease ($HBB^{Glu26Lys}$) is the most widely reported hemoglobin disorders after HbS. HbE is due to a mutation in which lysine is exchanged for glutamic acid in the Beta chain of hemoglobin

P. Pareek et al. (Eds.): IC4S 2023, LNICST 536, pp. 84–94, 2024.
https://doi.org/10.1007/978-3-031-48888-7_7

at the 26th position. SCD affects a sizable portion of population in India residing in the area which is spread throughout the Central Part of India from Odisha to Maharashtra and Gujarat. Sickle Hemoglobin (HbS) in the central belt of India, Hemoglobin E (HbE) in West Bengal and Northeastern States and Hemoglobin D (HbD_{Punjab}) in Northwest Parts of India are the three main genotypes of this disorder that are frequently observed in our country [1].

The most prevalent blood condition anemia is brought on by a deficiency of RBCs which makes it difficult for the body to get adequate oxygen. Acute anemia is caused by a sharp fall in RBC, whereas chronic anemia is caused by a gradual decline in RBC, and it frequently co-occurs with inflammatory illnesses. RBCs aren't formed as they should be in people with sickle cell disease. RBCs resemble round or spherical discs in normal human beings whereas they resemble a crescent moon, or an old farming tool called a sickle in SCD. It is a hereditary hemoglobin disorder as represented in Fig. 2.

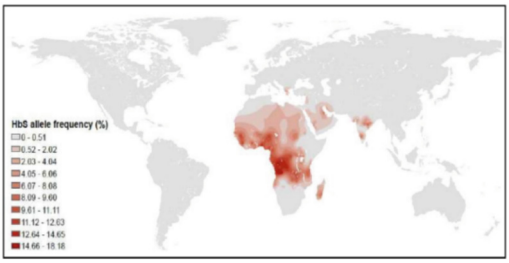

Fig. 1. Map depicting the global prevalence of the HbS allele [2]

Typically, SCA (Sickle Cell Anemia) symptoms and signs begin appearing around five months of age. Sickle cells quickly disintegrated and died, leaving just a small number of RBC in the circulation. The life of normal RBC typically lasts for about four months before they require replacement with new cells, while sickle RBCs often degrade in about two to three weeks, causing a lack of RBC.

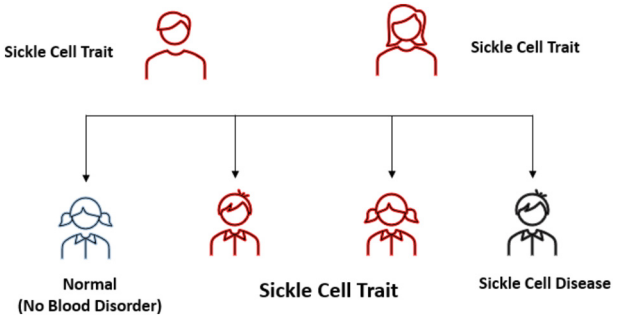

Fig. 2. Inheritance pattern of sickle cell hemoglobin gene from parents to the offspring

Individuals having sickle cell disease report edema, frequent infections, eye problems, slow growth and delayed puberty. As discussed, Hemoglobin S is an abnormal hemoglobin type that contributes to SCA. When both the parents pass the recessive sickle cell gene to their child, the child gets affected and develops the sickle cell homozygote phenotype.

SCD cannot be cured; however, it can be managed to lessen the symptoms and avoid complications. Hence, it is vital that these patients be monitored clinically and hemato-biochemical investigations on regular basis. Based on the steady state and crisis data patients can be monitored successfully [3, 4]. By analyzing the data using machine learning techniques the clinical and physiological state of the patients can be predicted well in advance to avoid any future complicacies. Keeping in view of this the present study was addressed. The data of SCD cases included in this study has been derived from a patient database of Odisha. Various tools of machine learning have been leveraged to predict and assess the health status of the patients those have been suffering from sickle cell anemia disease in Odisha.

2 Literature Survey

Sen et. al. (2021) took various microscopic blood samples and used techniques such as image processing and machine learning to make the process of detecting sickle cells automatic and have classified the RBC thus detected into three shape-based categories: circular, elongated sickle cell shaped and others, they are then preprocessed, and thresholding technique called Otsu is applied for segmentation [5].

Petrović et. al. (2020) used the smear from peripheral blood to observe the images of red blood cells and segmented the image by preprocessing and used machine learning techniques to classify their morphology [6].

In a case study of Nigeria Nkpordee (2022) have used different trend models of time series and statistics for a six-year projection of SCD in Nigeria and how it will decline in the year ahead [7].

Patel et. al. (2021) has shown how early detection of sickle cells can help patients to identify their symptoms and help the patients to take medications and can take regular blood transfusion sessions along with pain relieving medications. Sometimes manual assessment might lead to false classification. Therefore, using data mining techniques including classification algorithm they have sought to identify the sickle cells in human body with high accuracy [8].

Yang (2018) and Yeruva (2021) have employed machine learning algorithms to predictably understand the timing/situation of hospital re-admissions in SCD. In their research paper they have described how they partitioned their patients into groups for testing and training. The cases of unplanned treatment in the hospitals admissions were categorized for testing and training dataset where they applied machine learning algorithms. The prediction was then later assessed using various prediction algorithms such as specificity, sensitivity, and C-Statistic [9, 10].

Dean (2019) has used Multinomial Logistic Regression where they analyzed the pain scores of forty patients, and they devised a model of machine learning to predict the pain scores of a SCD patient with promising results [11]. Using proper optimization

techniques, machine learning algorithms, and statistics, it can be predicted whether the number of patients suffering from SCD will decrease with the use of a proper data set and patients as input.

A low cost, cost effective easy to use sickle cell screening device is proposed to be used in developing countries as elucidated by Wing (2019) [12] can detect hemoglobin non-invasively.

Stone (2021) [13] in this case report demonstrated the severity of a delayed hemolytic transfusion reaction caused by anti-Fy3 in a SCD patient having red cell exchanging before hematological precursor cell harvest for gene therapy.

Ranjana (2020) [14] used a automatic categorization of the SCA system explored in this study. In the beginning, the original images are pre-processed using the median filter. The Grey Level Co-occurring Model (GLCM) and Haralick characteristics are then retrieved. Finally, for prediction, the random forest (RF) predictor is used. Using an RF classifier, the SCA system achieves a classification accuracy of 95%.

Patgiri [15] demonstrated a hybrid segmentation procedure that combines two segmentation approaches, notably fuzzy C-means segmentation with adaptive (local) thresholding. In this study, four distinct adaptive thresholding approaches are used with fuzzy C-means. The main axis, and secondary axis, aspect ratio, surface dimension, circumference, dimension factor (metric value), eccentricity, and solidity of each cell in the sample blood smear were retrieved for this analysis. These eight characteristics are used to train and test the classifiers. For categorization, two supervised classifiers, namely the Nave Bayes classifier and the K-nearest neighbor classifier, were exhibited on a dataset of ten image data samples, and the evaluated results for all of the hybrid combinations were compared.

Even though a lot of techniques have been used or the prediction or image segmentation in various machine learning dataset the complicated clinical symptoms of SCD have not been addressed fully till date. So, it is necessary to predict the outcome for the year ahead and come up with some solution that will prevent the patients going through the tedious process of regular blood transfusions and doctor visits. Although these methods yield the best results but considering the complex clinical manifestations of symptoms from patients to patients implementing those methods has been challenging so far.

3 Linear Regression

Linear Regression is a widely known and recognized algorithm and is categorized under supervised learning technique. When a set of independent variables is given, the logistic regression is used for carrying out the prediction of dependent variables categorically, such that output result can be a categorical or discrete value. As it is an analysis of independent variable, and it can be represented in Eq. 1:

$$y = c0 + c1x + e \tag{1}$$

wherein y has been assumed to be a dependent variable and x to be an independent variable, the variable c0 is a constant term and an intercept of the regression line on the vertical axis and c1 is the regression coefficient that lies on the slope of regression line and e can be a random error as shown in Fig. 3.

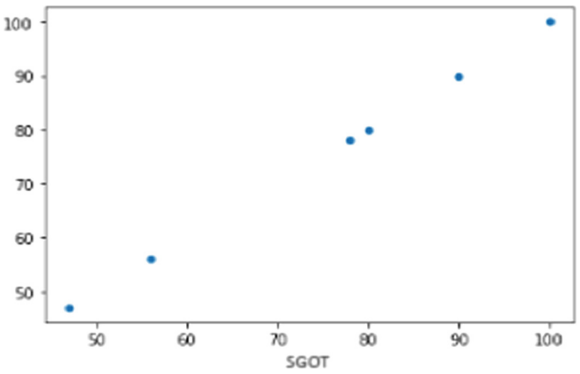

Fig. 3. Graph of linear regression

As specified earlier, the goal is to find out the best possible values for c0 and c1, the objective should be to minimize the error between the predicted value and the actual value as shown in Eqs. 2 and 3.

$$Minimize 1/n \sum_{i=1}^{n} (pred(i) - y(i))^2 \tag{2}$$

$$Q 1/n \sum_{i=1}^{n} (pred(i) - y(i))^2 \tag{3}$$

This function as mentioned above is aimed at minimizing the error values among the actual and the predicted values. Here the error difference is squared, added up across all the data points and then divided by the total number of data points. The result obtained (Q) is the average squared error across all data points. Hence the above cost function is also referred to as Mean Squared Error (MSE) function. With the MSE function the values of c0 and c1 are changed so that MSE settles at the minima.

Figure 4 depicts the process of fitting a linear regression model. Import data as an input, fit an optimization technique and a cost function for performance, verify its quality, change it to increase quality, and then find an output for the workflow.

Stochastic Gradient Descent (SGD) is a form of gradient descent variant used to optimize machine learning models. Only one random training example is used in this variant to calculate the gradient and update the parameters at each iteration. This algorithm is useful when the optimal points are not found by equating the slope of the function to zero (0). Linear regression on the other hand has the sum of squared residuals mentally mapped as the function "y" and the weight vector as "x" in the parabola above.

4 Proposed Work

The proposed approach extracts sickle cell data from a hospital in the western part of Odisha. The extracted dataset was then analyzed and usedS it as input for our machine learning model. So, first, we preprocessed and cleaned the data, and then we fed it into appropriate models for training and testing. Following the visualization and train-test

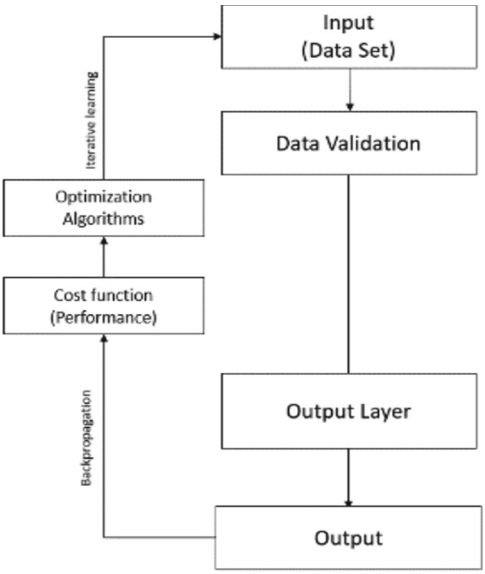

Fig. 4. Flowchart of the working of linear regression while using a dataset.

split, we selected an acceptable model (here, linear regression) for our planned work. We arrived at a proper conclusion after obtaining the accuracy and proper graphs (as indicated in the Graphs and Results section) (Fig. 5).

By using these statistical methods, we can classify, predict and find an optimal model that can help us identify the health / clinical status of people affected from sickle cell anemia or who have less amount of hemoglobin produced in their body.

Pseudocode
Input: *Patience dataset*
Pre-processing of data
Divide data into Train and Test with 80% and 20% respectively.
For each data in dataset
Linear regression (Train, Test data)
Perform gradient descent.
Predict test result.
Output: *predicted percentage, Correlation matrix*

5 Experimental Results

The proposed work as presented under section III has been analyzed using stochastic gradient descent analysis in conjunction with three different machine learning algorithm such as linear regression, decision tree classifier and support vector machine. The data

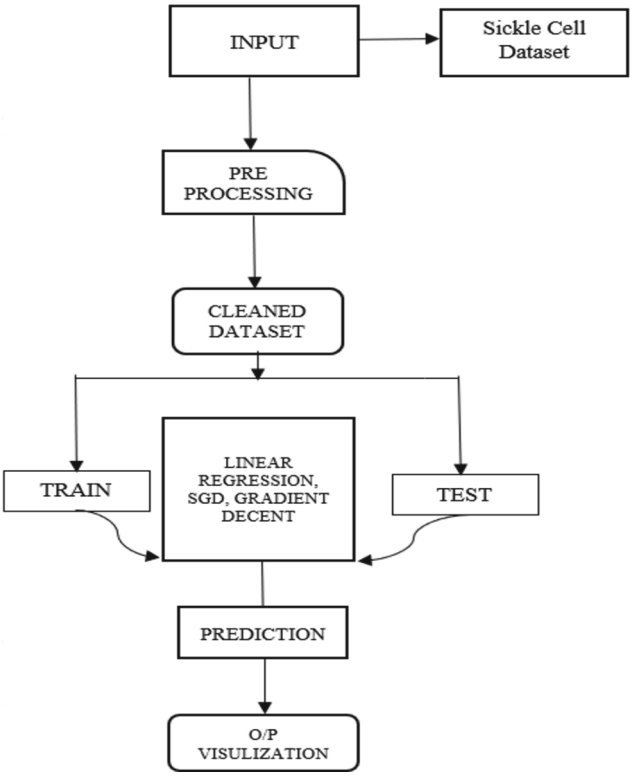

Fig. 5. Flowchart describing the methodology that was used in this study.

has been collected from the western part of Odisha state, highly affected with SCD. The data has been collected under six different categories such as WBC, RBC, HGB, BIT, LDH and RETICS%. All the data has been converted into it's per unit level except RETICS which is in percentage.

The data has been processed for redundancy analysis with a new value of 0.17. The scattered plot analysis has been carried out for two target data namely LDH, RETICS% with all the other four parameter as predictor. The effect of LDH and RETICS categorically analyzed with HGB. Out of the total dataset 80% data has been used for training and 20% has been used for testing the model.

Table 1 shows the statistical data of three different analyses of three different ML algorithms such as Decision Tree, Linear Regression and Support Vector Machine. It is found that the linear regression is having lowest RMSE of 3.60 and R^2 error is –0.39. However, the mean average error is 2.72; this could be due to similar type of data in the available training dataset. Further analysis has been carried out with Linear Regression with Gradient Descent.

Table 1. Statistical analysis of ML algorithms

	RMSE	MSE	R^2 Error	MAE
DT	4.32	18.70	–0.93	2.41
LR	3.60	13.03	–0.39	2.72
SVM	5.32	28.30	–1.92	2.56

Figure 6 shows the regression analysis of RETICS vs HGB where most of the data possesses negative slope characteristics which means that with increasing HGB content there is a decrease in the RETICS. In most of the cases the HGB content carries in between 8.5 to 11, which corresponds to a decrease in 17% of RETICS%

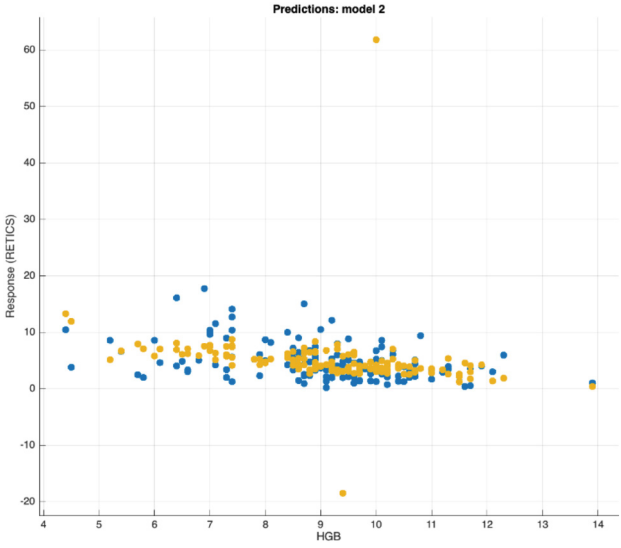

Fig. 6. Regression analysis of RETICS vs HGB

Similarly, Fig. 7 represents the statistical graphical analysis of LDH vs HGB. However, with the same range of HGB (refer Fig. 6) the LDH content varies between 385–460.

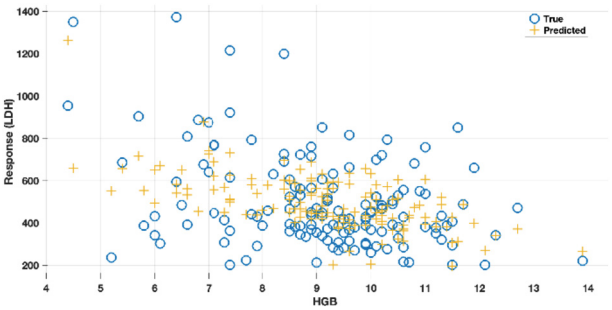

Fig. 7. Regression analysis of LDH vs HGB

Table 2 shows the analysis of different ML algorithms for LDH vs HGB. Figure 8 shows the heat map based on autocorrelation function where all the diagonal elements have a magnitude 1 per unit.

Table 2. Statistical analysis of ML algorithm

	RMSE	MSE	R^2 Error	MAE
DT	383.04	1.46 x 10^5	–1.08	180.6
LR	274.32	75251	–0.07	181.41
SVM	286.16	81885	–0.165	155.05

Table 3. Correlation statistical analysis with Hb

Sr. No	Parameter	Magnitude	Remarks
1	RETICS%	–0.39248	NEGATIVE
2	LDH	–0.37992	NEGATIVE
3	BIT	–0.137320	NEGATIVE
4	RBC	0.300613	POSITIVE
5	WBC	–0.09364	NEGATIVE

Figure 8 this correlation depicts that the patient does not have enough hemoglobin produced in this body as the formation of LDH is high in their body hence they cannot carry enough oxygen to supply throughout their body.

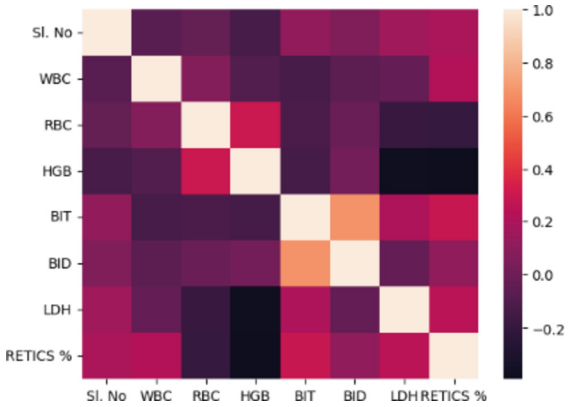

Fig. 8. Heatmap of the parameters.

6 Conclusion and Future Work

From the present study with the data of the patients who have sickle cell disorder (HbSS) and using the ML models it can be predicted that the patients having low hemoglobin might face many clinical symptoms due to the formation of high level of LDH, more numbers of WBC and reticulocyte counts. Due to the low hemoglobin in their body they are unable to meet the oxygen demand of the body and subjected to deoxygenated state and leads to high amount of lactic acid and higher count of reticulocyte or premature RBCs in their circulation. Hence patients must go through regular blood transfusions and hemoglobin tests. The traditional method of measurement of hemoglobin is accurate, infants and adults are hesitant to use it since it is painful, and regular blood extraction makes them uncomfortable. As a result, introducing a non-invasive way will be beneficial to determine their hemoglobin level any place and without any pain. Till then regular management of the patients with the clinical and hematological data set using machine learning techniques will be of great importance.

References

1. Chhotray, G.P., Dash, B.P., Ranjit, M.: Spectrum of Hemoglobinopathies in Orissa, India (2004). https://doi.org/10.1081/hem-120034244
2. Piel, F.B., et al.: Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis (2010). https://doi.org/10.1038/ncomms1104
3. Serjeant, G.R.: One hundred years of sickle cell disease (2010). https://doi.org/10.1111/j.1365-2141.2010.08419.x
4. Meher, S., et al.: Haptoglobin Genotypes Associated with Vaso-Occlusive Crisis in Sickle Cell Anemia Patients of Eastern India (2021). https://doi.org/10.1080/03630269.2020.1801459
5. Sen, B., Ganesh, A., Bhan, A., Dixit, S., Goyal, A.: Machine learning based diagnosis and classification of sickle cell anemia in human RBC. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, pp. 753–758 (2021). https://doi.org/10.1109/ICICV50876.2021.9388610

6. Petrović, N., Moyà-Alcover, G., Jaume-i-Capó, A., González-Hidalgo, M.: Sickle-cell disease diagnosis support selecting the most appropriate machine learning method: towards a general and interpretable approach for cell morphology analysis from microscopy images (2020). https://doi.org/10.1016/j.compbiomed.2020.104027
7. Nkpordee, L., Wonu, N.: Statistical modelling of genetic disorder in Nigeria: a study of sickle cell disease. Faculty Nat. Appl. Sci. J. Sci. Innov. 3(2), 10–19 (2022). https://www.fnasjournals.com/index.php/FNAS-JSI/article/view/27
8. Patel, A., et al.: Machine-learning algorithms for predicting hospital re-admissions in sickle cell disease. Brit. J. Haematol. **192**(1), 158–170. Wiley (2020). https://doi.org/10.1111/bjh.17107
9. Yang, F., Banerjee, T., Narine, K., Shah, N.: Improving pain management in patients with sickle cell disease from physiological measures using machine learning techniques. In: Smart Health, vols. 7–8, pp. 48–59. Elsevier BV. (2018). https://doi.org/10.1016/j.smhl.2018.01.002
10. Yeruva, S., Gowtham, B.P., Chandana, Y.H., Varalakshmi, M.S., Jain, S.: Prediction of anemia disease using classification methods. In: Machine Learning Technologies and Applications, pp. 1–11. Springer, Singapore (2021). https://doi.org/10.1007/978-981-33-4046-6_1
11. Dean, C.L., et al.: Challenges in the treatment and prevention of delayed hemolytic transfusion reactions with hyperhemolysis in sickle cell disease patients. Transfusion **59**(5), 1698–1705. Wiley (2021). https://doi.org/10.1111/trf.15227
12. Wing, J., et al.: A low-cost, point-of-care sickle cell anemia screening device for use in low and middle-income countries. In: 2019 IEEE Global Humanitarian Technology Conference (GHTC), Seattle, WA, USA, pp. 1–4 (2019). https://doi.org/10.1109/GHTC46095.2019.9033017
13. Stone, E.F., et al.: Severe delayed hemolytic transfusion reaction due to anti-Fy3 in a patient with sickle cell disease undergoing red cell exchange prior to hematopoietic progenitor cell collection for gene therapy (2020). https://doi.org/10.3324/haematol.2020.253229
14. Ranjana, S., Manimegala, R., Priya, K.: Automatic classification of sickle cell anemia using random forest classifier. In: Proceedings of the European Conference on Medical Advances, LNCS, vol. 9999, p. 2020. Springer, Heidelberg (2020)
15. Patgiri, C., Ganguly, A.: Adaptive thresholding technique based classification of red blood cell and sickle cell using Naïve Bayes Classifier and K-nearest neighbor classifier (2021).https://doi.org/10.1016/j.bspc.2021.102745