EasyChair Preprint

№ 3164

Prediction of Anemia Disease Using Classification Methods

Bathula Pavan Gowtham, Yendluri Hari Chandana,
Sagar Yeruva, M. Sharada Varalakshmi,
P. E. S. N. Krishna Prasad, Suman Jain,
Allam Ravi Kumar Reddy, Saroja Kondaveeti and Padma Gunda

April 13, 2020

Prediction of Anemia Disease Using Classification Methods

B Pavan Gowtham (M Tech)[1], Hari Chandana (B Tech)[1], Dr. Sagar Yeruva[1]
[1] *Department of CSE, VNRVJIET, Hyderabad, India.*
Dr. M. Sharada Varalakshmi[2]
[2] *Department of CSE, St. Peter's Engineering College, Hyderabad, Telangana, India.*
Dr. P. E. S. N. Krishna Prasad [3]
[3] *Department of CSE, SV College of Engineering, Tirupathi, Andhra Pradesh, India.*
Dr. Suman Jain (Chief Medical Research Officer and Secretary) [4], Allam Ravi Kumar Reddy (Data Manager) [4], Dr. Saroja Kondaveeti (Medical Officer) [4], Dr. Padma Gunda (Research Scientist) [4]
[4]*Thalassemia and Sickle Cell Society, Rajendra Nagar, Hyderabad, Telangana, India.*
[1] pavangowtham2495@gmail.com,
[1] hcyendluri@gmail.com,
[1] sagar_y@vnrvjiet.in
[2] sharada.mangipudi07@gmail.com
[3]surya125@gmail.com
[4]sumanjaindr@gmail.com,
[4]allamravikumar@gmail.com,
[4] drsarojakondaveeti@gmail.com,
[4] padma.genetics@gmail.com

Abstract

Sickle cell is a hematological disorder (hematological is a study of blood in health and diseases) which may leads Organ damage, heart strokes and serious complications. It may also reduce the human lifespan. Most of the sickle cells are served in newborn babies. It was initially thought to be a particular feature of tribal peoples, but it has now been found in all populations. Sickle cell Symptoms are observed in human beings as episodes of pain (crisis), painful swelling of hands, feet and Vision problems. Detecting sickle cell as early as possible could help the patients to identify their symptoms and can support to take the medications using Antibiotics, Vitamins, Blood transfusion and pain relieving medicines etc. The manual assessment, classification and Counting of cells require for an intense spending of time and it may lead to wrong classification and counting, since red blood cells are millions in one smear. By using various data mining techniques like classification algorithms, we can identify sickle cells in the human body effectively with high accuracy.

The proposed method overcomes the drawbacks of manual assessment by introducing robust and effective classification algorithms to classify the Sickle Cell Anemia (SCA) in blood cells into three classes namely: Normal (N), Sickle Cells(S) and Thalassemia (T). In this paper we also present the accuracy levels of various classification algorithms on the dataset that we gathered from Thalassemia and Sickle Cell Society (TSCS) located at Rajendra Nagar, Hyderabad, India

Keywords: *Sickle Cell (SC), Sickle Cell Disease (SCD), Thalassemia, Sickle Cell Anemia (SCA), Anemia, classification.*

I Introduction

A Normal Blood flows through a small circular shape which Carries oxygen to organs of human body parts which is circular in shape and life span of each cell is approximately 120 days and a new blood cell is generated for every 120 days[1]. Sickle Cell Anemia is a one kind of abnormal blood disease which affects normal hemoglobin within the red blood cells and it is called as Sickle cell Disease or a normal Sickle Cell. Shape of the sickle cell is Disc shape which is Sticky and rigid, which causes stoppage of blood flow in the human body. [2] Sickle cell life span is 10 to 20 days. Due to the presence of Sickle cell in hemoglobin, it may cause severe episodes of pain, death of tissue and serious complications, in some cases it may lead to death [3].

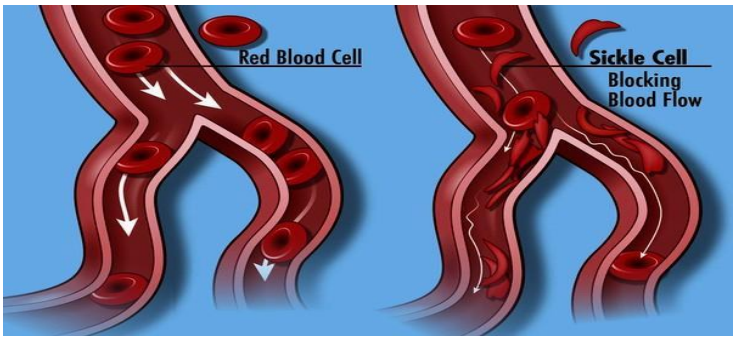


Figure 1 Normal Blood Flow and Sickle cell Blood Flow

In above Figure 1, we can clearly see that the normal blood flow passing oxygen to all parts without stoppage of Blood, where sickle cell shape is sticky and due to that blood flow will stop at any stage. Due to stoppage of blood flow may cause severe body pains and heart strokes etc. Sickle cell was observed in the black population, later it has been seen in people of other ethnic group, which include individuals from parts of the Middle East, Central India, Mediterranean Sea, especially in Italy and Greece [4]. Treatment for Sickle cell Disease by taking early medications like Antibiotics, Blood Transfusion, Bone marrow. Most of Doctor recommends blood Transfusion for at least 1-2 months and patient can take antibiotics to manage complications, including chronic pain. SCD is a rare blood Disorder in the human body but now it has been observed very quietly in newborn babies' hemoglobin. SCD are crescent-shaped cells which are stiff and sticky and interact with other cells. If one organ is affected slowly it will be spreading throughout the body which may lead to death. To prevent this kind of disease, early medication required as soon as possible to avoid serious complications (Blood Transfusion).

When a person has hemoglobin S which is caused by two abnormal genes then the person has Sickle Cell Disease.

1.1 SYMPTOMS AND SIGNS OF SICKLE CELL ANEMIA [5]

1) **Episodes of pain**: It is one of the major symptoms of sickle cell where RBC blocks the blood flow through tiny blood vessels joints, abdomen and chest. Chronic

pain has been observed in adults and adolescents which results in bones, ulcers, and joint damage.

2) **Painful swelling of hands and feet**: Due to sickle cell presence in bodies, RBC blocks the blood flow through hands and feet, it may cause swelling in hands and feet.

3) **Frequent infections**: Organ damage (fights infection i.e. spleen), may cause due to sickle cells in bodies. Doctors commonly suggest vaccinations and antibiotics that can prevent life-threatening infections, such as pneumonia.

4) **Delayed growth**: If RBCs are getting reduced in the human body due to lack of oxygen, which may cause delayed growth.

5) **Vision problems**: RBC blocks the blood flow through tiny blood vessels due to sickle cells. These cells do not supply blood to the eyes which may lead to damage of retina. It causes blur vision problems.

1.2 SICKLE CELL COMPLICATIONS [5]

1) **Heart Stroke**: Heart strokes may occur if sickle cells block the blood flow in brain. Symptoms of strokes include weakness or numbness of arms and legs, continuously speaking problem, loss of consciousness.
2) **Acute chest syndrome**: It causes chest pain, fever and difficulty in breathing. It is a very serious complication, which may also lead to life threatening. It requires emergency medical treatment by taking antibiotics and any other treatments.
3) **Pulmonary Hypertension**: Patients with sickle cell anemia can have high blood pressure in their lungs (pulmonary hypertension). This complication affects adults rather than children. Symptoms of this condition may cause breathing problems and fatigue.
4) **Leg Ulcers**: SCA (sickle cell anemia) can cause open sores, in which ulcers may occur in legs.
5) **Gallstones**. The breakdown of RBC may produce a substance i.e. bilirubin. If bilirubin is having a high level in the body which leads to gallstones.

1.3 STATISTICAL INFORMATION [6]

In Indian Scenario [6]

1) First described in the Nilgiri Hills of northern Tamil Nadu in 1952.

2) The sickle cell gene is now known to be widespread among people of the Deccan plateau of central India with a smaller focus in the north of Kerala and Tamilnadu.

In worldwide Scenario [6]

1) Sickle cells was firstly an unknown disease but now is been spreading over worldwide and it is observed particularly in Spanish Speaking regions in the Western Hemisphere (The Caribbean, South America and Central America), Saudi Arabia, India, Mediterranean countries include(Greece, Italy, Turkey), sub-Saharan Africa.

2) Sickle cell deaths are observed more in African-American, with children younger than 4 years of age with sickle cell disease fell by 42% from 1999 through 2002. With the help of vaccination the drop over sickle cell is observed in 2000 which protects against invasive pneumococcal disease.
3) During 1990-1994 Sickle Cell Disease Identified by Newborn Screening among mortality children in California, Illinois, and New York.
4) By the end of 1995, the mortality rate is 1.5 per 100 in African-American children with SCD in California and Illinois. The mortality rate in African-American or black infants born during this period in California and Illinois is 2.0 per 100 Black or African-American or Black.
5) SCD is one of the major public health concerns where the average of 75000 Hospitalization due to SCD in the US, approximately cost $475 millions.

II Literature Review

In this section we present various approaches / methods available in the existing sources that are mainly helpful in the identification of sickle cell disease. Various scientists / researchers have made their efforts in the progress of identification of sickle cell disease in the early stage of life with good accuracy levels.

Classification Red Blood Cells Using Support Vector Machine [7]
Findings:

In this paper, image processing techniques that use the optimization segmentation and mean filter play an important role in obtaining the geometric, texture and color features related to RBC images by using a photo imaging microscope. The support vector machine, which is an advanced kernel-based technique, is used to classify RBC data as either normal or abnormal, this method gives accuracy rates in the form of validation measure of sensitivity, specificity and Kappa to be 100%, 0.998% and 0.9944 respectively.

Anemia cells detection based on Shape Signature using NN [8]
Findings:

Elsalamony et al [8] has used a method for detecting anemia kind of disease such as sickle cells and elliptocytosis with their geometrical shapes signatures method.
They use 30 colorful microscopic images and achieved 100% based on the three Neural Networks for the anemia kind of disease**.**

Detection of anemia disease in human red blood cells using cell signature, Neural Networks and SVM [9]
Findings:

Identifying the sickle cell, Burr cells, Elliptocytosis based on shape signatures. They use 45 colorful microscopic images in 15 samples by using Circular Hough Transforms watershed segmentation and some of the morphological methods to enhance image. The

Support Vector Machine (SVM), back propagation (BP) and Self- Organizing Map (SOM) neural networks used to identify anemia kind of disease better accuracy.

Classification of three types of red blood cells in peripheral blood smear based on morphology [10]
Findings:

In this paper, Elliptocytes, Discocytes and Echinocytes, which are three known R.B.Cs, are classified in a peripheral blood smear based on morphological methods. For this reason, a simple statistical analysis of distances of each edge pixel from the mass center is employed. This method is 98.63% successful for Elliptocyte recognition, 96.7% successful for the normal Discocyte recognition and 95.36% successful for Echinocyte recognition.
The process is

- Preprocessing- converting image to grayscale for accurate, segmentation-threshold (converting image to binary image),
- Edge detection -After microscopic image enhancement, thresholding, filling and morphological closing, in this step we can extract the edge of cells from these binary images by a gradient matrix.

Data mining technique using WEKA classification for Sickle Cell Disease [11]:
Findings:

In this paper, Elsalamony used two classification techniques like J48 and Random tree for prediction of sickle cell disease which is highly disease highly affected to tribal zones of Gujarat and after that the author compared J48 and Random tree classification techniques for the mining process. They have used the WEKA tool for this prediction process, which is an open source tool.

III About dataset

As part of this project [A collaborative Research project granted by Jawaharlal Nehru Technological University, Hyderabad, TEQIP-III (funding agency) for Rs.3,00,000/- for the duration of 1 year (August, 2019 to July, 2020)], we had executed a Memorandum of Understanding (MoU) between our Institute, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology (VNRVJIET) and Thalassemia and Sickle Cell Society (TSCS) (MoU dated 15 Oct 2019). The details of TSCS can be found from https://www.tscsindia.org/.

The MoU aimed with the following objectives:

1: Data Sharing and

2: Technology Transfer

We have received a dataset of 1387 records of patients who have approached TSCS for the diagnosis purpose. These records are shared to us as part of MoU and the records are

preprocessed to maintain confidentiality, anonymity that meets data privacy of patients. This data of the patients gathered during August, 2017 to August, 2019.

After preprocessing we have identified 13 attributes along with class label (Diagnosis of Blood sample) these attributes are summarized as follows

IV Implementation

The model architecture includes three stages include:

1. Dataset preparation.
2. Analyzing dataset and splitting dataset
3. Model selection

4.1 Dataset preparation:

As we mentioned above about the dataset collected from Thalassemia and Sickle Cell Society is about 1387 patients with 13 parameters which include: 1) **AGE ,** 2) **HB**-Hemoglobin, 3) **HCT**-Hematocrit, 4) **RDW** - RBC Distribution Width, 5) **MCV** -Mean Corpuscular Volume, 6) **MCH** -Mean Corpuscular Hemoglobin, 7) **MCHC** - Mean Corpuscular Hemoglobin Concentration 8) **RBC** : Red Blood Cell, 9) **RETIC** – Reticulocytes, 10) **HBF** - Fetal Hemoglobin , 11) **HBAo** 12) **HBA2** 13) **Diagnosis** - Diagnosis is output variable, which we need to predict based on set of features (inputs) either it is Normal cell, Sickle cell or Thalassemia cells.

4.2 Analyzing dataset and splitting dataset:

Thalassemia society have labeled each patient data record (i.e. Diagnosis) with AS , AT, BT, EA, EE, ET, NN, SC, SS, ST, TM,TT . 12 labeled Diagnosis is converted into three main groups

1) **NN** = N (Normal Cells)

2) **AS, SC, SS, ST** = S (Sickle Cells)

3) **AT, BT, EA, EE, ET, TM, TT** = T (Thalassemia Cells)

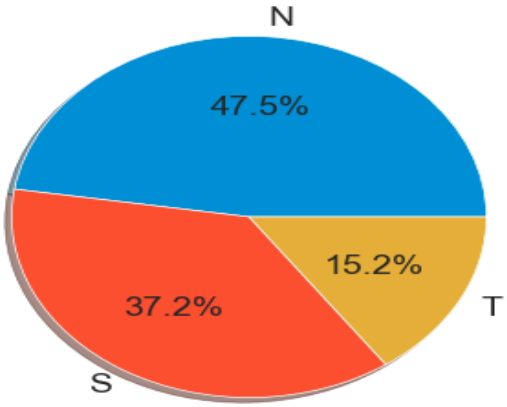


Figure 2: Dataset collected from TSCS

Label Encoder is the part of SciKit Learn library in Python and used to convert categorical data, or text data, into numbers, which our predictive models we can understand easily.

The dataset is split into training data and test data (80% and 20%.) The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

4.3 Model Selection:

Model selection is an important part as we are applying Machine learning algorithms for predicting the best outcome or result.

Supervised technique is used to train on a set of input and output pairs and learn the model for correlations between inputs and outputs. Supervised learning problems are grouped into two problems:

1) **Regression analysis:** when the target or output variable is real and continuous values**.**

2) **Classification:** Problems for filtering the data which is not required.

We have a dataset with 13 Independent variables (parameters or features) and one Dependent variable is target variable or output variable which predicts whether patients are having either **N** (Normal) or **S** (sickle cell) or **T** (Thalassemia). In this project we have used various classification algorithms for the diagnosis purpose and are elaborated as follows:

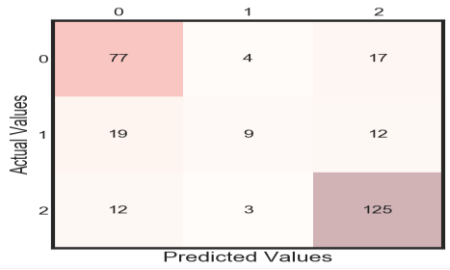
We test this dataset using data mining techniques like classification method [13] for the detection of the target classes described as above.

Classifiers: It is a supervised learning technique which allows computers to learn from the data. Input given to it and then uses this learning to classify new observations. We have tested the dataset with the following classifiers which include:

1) SVM 2) KNN 3) Logistic Regression 4) Decision Tree 5) Random Forest

SVM classifier: SVM is a supervised learning technique which evaluates the statistics used for analysis of regression and classification.SVM algorithm provides best possible Decision Boundary, so we can categorize data points easily. It selects extreme points that support the hyper plane imagination, in such cases called as vectors of support and algorithms used for machine learning are called Vector Support Machines.

Results obtained using SVM: The following is the accuracy and other attributes defines the performance of this method on this dataset.

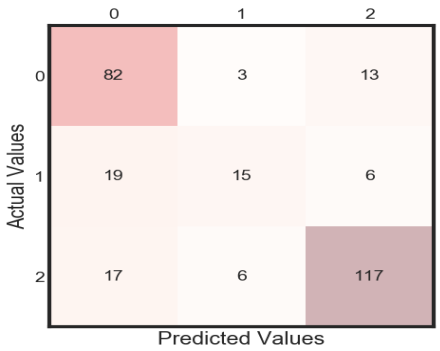


	precision	recall	f1-score	support
N	0.71	0.79	0.75	98
S	0.56	0.23	0.32	40
T	0.81	0.89	0.85	140
accuracy			0.76	278
macro avg	0.70	0.63	0.64	278
weighted avg	0.74	0.76	0.74	278

Figure 3: Results obtained using SVM classifier

KNN Classifier: K-Nearest Neighbors (**KNN**) is one of the simplest algorithms used for regression and classification problems. Based on similarity measures it classifies new data points.

Results obtained using KNN Classifier:

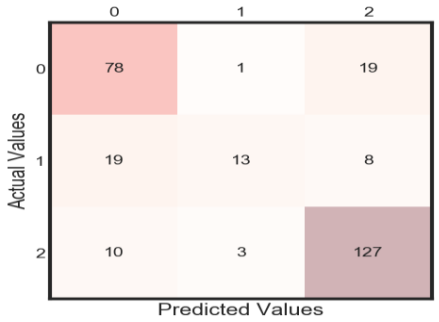


	precision	recall	f1-score	support
N	0.69	0.84	0.76	98
S	0.62	0.38	0.47	40
T	0.86	0.84	0.85	140
accuracy			0.77	278
macro avg	0.73	0.68	0.69	278
weighted avg	0.77	0.77	0.76	278

Figure 4: Results obtained using KNN Classifier

Logistic Regression Classifier: Logistic regression is a supervised **classification** algorithm. In a **classification** problem, the output variable (or Y), which accepts only discrete values for a given set of features (inputs or X), is called logistic **Regression**.

Results obtained using Logistic Regression Classifier:

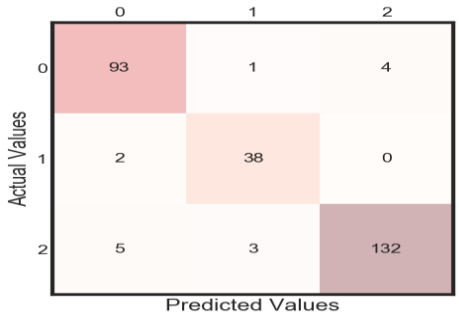


N	0.73	0.80	0.76	98
S	0.76	0.33	0.46	40
T	0.82	0.91	0.86	140
accuracy			0.78	278
macro avg	0.77	0.68	0.69	278
weighted avg	0.78	0.78	0.77	278

Figure 5: Results obtained using Logistic Regression Classifier

Decision Tree classifiers: A decision tree supports the decisions and their consequences in the form of a tree-structure model. Decision tree is used to define a structural approach that is most likely to meet an objective, especially in decision analysis.

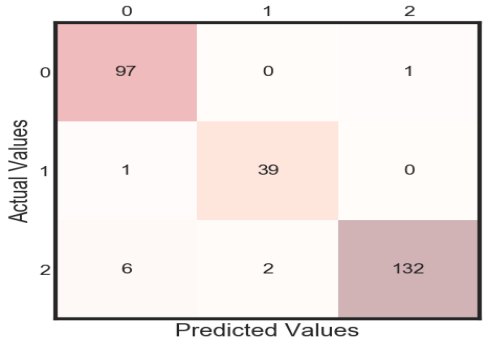
Results obtained using Decision Tree:



	precision	recall	f1-score	support
N	0.93	0.95	0.94	98
S	0.90	0.95	0.93	40
T	0.97	0.94	0.96	140
accuracy			0.95	278
macro avg	0.94	0.95	0.94	278
weighted avg	0.95	0.95	0.95	278

Figure 6: Results obtained using Decision Tree

Random Forest classifiers: A **random forest** is a meta estimator that allows a number of decision **tree classifiers** on various sub-samples of the dataset and uses average for improving the predictive accuracy and Avoiding over fitting.



	precision	recall	f1-score	support
N	0.93	0.99	0.96	98
S	0.95	0.97	0.96	40
T	0.99	0.94	0.97	140
accuracy			0.96	278
macro avg	0.96	0.97	0.96	278
weighted avg	0.97	0.96	0.96	278

Figure 7: Results obtained using Random Forest algorithm

```
from sklearn.preprocessing import LabelEncoder
dataset = pd.read_excel("C:/Users/pavan/sickle_cells_Anemia/SCA5.xlsx")
#14 labels are there present in the dataset
X = dataset.iloc[:, 1:13].values
Y = dataset.iloc[:, 13:].values
#Printing the Dataset Dimensions
print("sca dimensions : {}".format(dataset.shape))
#Checking for NULL values in the Dataset
a = dataset.isnull().sum()
b = dataset.isna().sum()
dataset['Diagnosis'] = dataset['Diagnosis'].map({'N': 0, 'S': 1, 'T': 2})
print(dataset.groupby('Diagnosis').size())
labelencoder_Y = LabelEncoder()
Y = labelencoder_Y.fit_transform(Y)
#Splitting the dataset for test and train
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20, random_state=0)
print(X_test.shape)
models_list = []
from sklearn.svm import SVC
import time
models_list.append(('DTC', DecisionTreeClassifier()))
models_list.append(('SVM', SVC()))
models_list.append(('RF', RandomForestClassifier()))
models_list.append(('KNN', KNeighborsClassifier()))
models_list.append(('LG', LogisticRegression()))
```

Figure 8: A sample code

V. RESULTS

Analysis of results:

Experimental results have done through three stages:

1. One is the detection process for the normal cells, sickle cells, Thalassemia cells
2. Training the dataset and making prediction using testing data
3. Calculating accuracy based on Test Predictions, whether patient is having normal cells, sickle cells or Thalassemia.

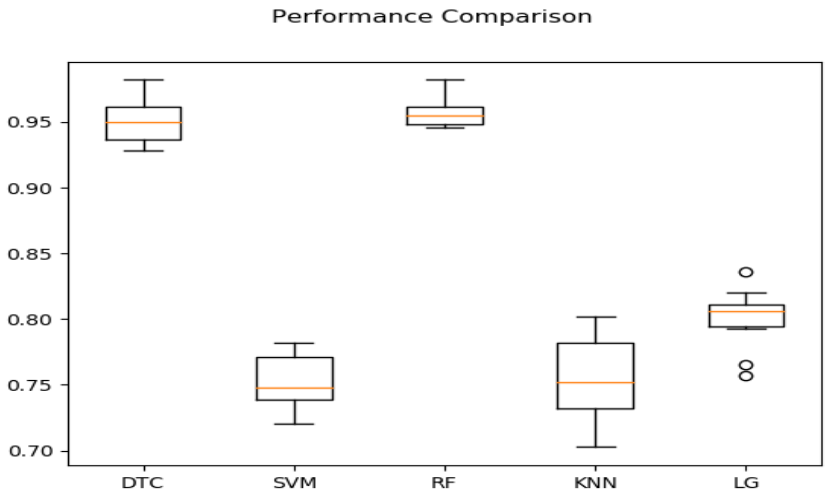
$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total number of Predictions}}$$

Prediction using classifiers:

SNO	Name of the Classification Algorithm	Result of prediction accuracy (%)	Dataset Size (No of observations X No of Parameters)
1	Support Vector Machine (SVM)	76	1387 X 13
2	K-Nearest Neighbor Algorithm (KNN)	77	1387 X 13
3	Logistic Regression (LG)	78	1387 X 13
4	Decision Tree Classifier (DTC)	95	1387 X 13
5	Random Forest Algorithm (RF)	96	1387 X 13

Table-1: Performance of various classification algorithms

Performance Comparison:



VI. Conclusion:

Sickle cell Disease is a haematological disorder, previously it was one of the particular features of tribal people but now it is spreading to the entire world in a rapid manner and needs immediate attention. This paper has presented a description about Sickle Cell Disease (SCD) and its history in the international scenarios and in the national scenario (in the context of India). The symptoms of the disease, signs, complications, and treatment for the disease are presented. We have also characterized the blood cells in normal people with that of sickle cell and Thalassemia patients under various blood parameters. This paper also presents the identification of sickle cell disease with higher accuracy using various classifiers and to develop good prediction models that help to reduce the time and efforts in pain management systems of sickle cell disease. The results shows that the classification algorithms like SVM, KNN, Logistic Regression, Decision Tree and Random Forest gives the accuracy of prediction like 76%, 77%, 78%, 95% and 98% respectively. Finally the results shown that the Random Forest algorithm gives the best accuracy of prediction in the identification of sickle cell from the blood samples.

VII. ACKNOWLEDGEMENTS:

1. JNTUH, TEQIP-III: We sincerely thank Jawaharlal Nehru Technological University, Hyderabad, Technical Education Quality Improvement Programme-III (JNTUH TEQIP-III) for the award of project (proceedings No: JNTUH/TEQIP-III/CRS/2019/CSE/04 Dated 22-07-2019) with an amount of Rs 3,00,00/- for the duration of 1Year (Aug 2019 to July 2020) as part of Collaborative Research Project Scheme. We also sincerely thank the reviewers and coordinator of JNTUH-TEQIP-III, Dr. Padmaja Rani, Professor, Department of CSE, JNTU Hyderabad for their constant support during the reviews for carrying out this project.

2. TSCS: We sincerely thank Thalassemia and Sickle cell society (TSCS) for accepting our request for Collaborative Research Project and their support for entering into MoU. We also thank the personal at TSCS named Mr. Chandrakant Agarwal, President-TSCS, Dr. Suman Jain, Chief Medical Research Officer and Secretary-TSCS, Mr. Allam Ravi Kumar Reddy, Data Manager-TSCS, Dr. Saroja Kondaveeti, Medical Officer-TSCS, Dr. Padma Gunda, Research Scientist-TSCS, Mr. Mohd Abdul Tufeeq Baig, Lab Incharge-TSCS, Mr. Bhargava Kalvakota, Data & Admin Officer-TSCS and Ch. Devasri, Data Entry Operator-TSCS who have helped us to understand the entire scenario of sickle cell patients, process of their work and their services to the society in the state of Telangana, India.

REFERENCES

[1] *https://en.wikipedia.org/wiki/Red_blood_cell.*

[2] *Sickle Cell Disease core concepts for emergency physician and nurse (https://slideplayer.com/slide/3762536/)*

[3] *https://www.mayoclinic.org/diseases-conditions/sickle-cell-anemia/symptoms causes/syc-20355876*

[4] *https://www.nhlbi.nih.gov/health-topics/sickle-cell-disease*

[5] *https://www.mayoclinic.org/diseases-conditions/sickle-cell-anemia/symptoms*

[6] *causes/syc-20355876*

[7] *https://www.cdc.gov/ncbddd/sicklecell/data.html*

[8] *Akrimi, Jameela Ali, et al. "Classification red blood cells using support vector machine." Proceedings of the 6th International Conference on Information Technology and Multimedia. IEEE, 2014.*

[9] *Elsalamony, Hany A. "Anaemia cells detection based on shape signature using neural networks." Measurement 104 (2017): 50-59.*

[10] *Elsalamony, Hany A. "Detection of anaemia disease in human red blood cells using cell signature, neural networks and SVM." Multimedia Tools and Applications 77.12 (2018): 15047-15074.*

[11] *Soltanzadeh, Ramin, and Hossein Rabbani. "Classification of three types of red blood cells in peripheral blood smear based on morphology." IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS. IEEE, 2010.*

[12] *Solanki, Ashok kumar Vijaysingh. "Data mining techniques using WEKA classification for sickle cell disease." International Journal of Computer Science and Information Technologies 5.4 (2014): 5857-5860.*

[13] *https://www.tscsindia.org/*

[14] *https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/*