(International Print/Online Journal)

SJIF IMPACT FACTOR: 5.565
PUBMED-National Library of Medicine ID-101739732

ISSN (Print): 2209-2870
ISSN (Online): 2209-2862

International Journal of Medical Science and Current Research (IJMSCR)
Available online at: www.ijmscr.com
Volume 5, Issue 3, Page No: 887-894
May-June 2022

IJMSCR

Insights Into The Biochemical Factors Associated With The Severity In Sickle Cell Patients

Dr. Sudama Rathore, Dr. Vandana Dewangan, Dr. Neha Rani Verma, Dr. Abhigyan Nath
Department of Biochemistry, Pt.J.N.M. Medical College, Raipur (CG)

***Corresponding Author:**
Dr. Abhigyan Nath
Department of Biochemistry, Pt.J.N.M. Medical College, Raipur (CG)

Type of Publication: Original Research Paper
Conflicts of Interest: Nil

Abstract

Introduction: Sickle cell Disease is the oldest genetic disorder known and studied by the scientific society. It has been more than an era passed since the disease has captured the attention of medicine. A huge amount of effort has been put into the study of various aspects of the disorder since then . But even after so much efforts a lot about the pathophysiology of the disease process still remains unexplored.

Methods: In the current study, an explainatory machine learning approach is implemented for gaining insights into the biochemical paramters in relation to severity of sickle cell cases (Steady state and Crisis state).

Results: We implemented partial least square discriminative analysis (PLSDA) and random forest (RF) algorithms for developing prediction models to discriminate between Steady state and Crisis patient groups and obtained an accuracy of 72.5 % (PLSDA) and 82.5 % (RF). HbF, MCHC and MHC are the three most important features for the prediction model and stongly associated with the descrimination between the two group of patients.

Conclusion: Sickle cell crisis is one of the most dreadful complications of the disease amounting to a significant burden of morbidity and mortality of the affected individuals. Incomplete understanding of the cellular mechanisms of the process has been the primary limitation for the medical fraternity in combating sickle cell crisis. The features: HbF, MCHC and MCH are the three most important variables for the model and are strongly associated with the descrimination between the two group of patients. The current work establishes the association of physicochemical parameters with severity of sickle cell condition, achieving an acceptable overall accuracy.

Keywords: Sickle cell, random forest, partial least square analysis, crisis state

Introduction

Sickle cell disease (SCD) is an inherited disorder of hemoglobin (Hb) caused by substitution of a single nucleotide in the β-chain of hemoglobin resulting in amino acid valine instead of glutamic acid [1]. This leads to point mutation which is responsible for modification in the properties of the hemoglobin chain, which gets polymerized in the deoxygenated state [2], changing normal, flexible biconcave disc shaped red blood cells (RBCs) into stiff, inflexible,fragile, sickle cell.

The rate of polymerization of Sickle cell hemoglobin (HbS) is directly related to important pathophysiology of hemolytic anemia and vaso-occlusion [3]. The inheritance of one sickle gene (βs-globin) results in carrier for SCD (haemoglobin AS), whereas the inheritance of two abnormal βs-globin genes leads to SCD (HbSS). Sickle cell anaemia is transmitted as an autosomal recessive trait [4]. In Chhattisgarh, Sickle cell anaemia is common in central & south part i.e. a few district like Sarguja, Raigarh, Jashpur, Mahasamund & Korea. In caste wise distribution of SCD in Chhattisgarh , the

disease is more prevalent in the OBC communities like Agharia, Kurmi, Kolta, Teli, Kumhar , ST communities of Halba, Gond, Binjhwar & SC communities of Ghasia, Gada & Mahar [5].

The vaso-occlusive crisis,which is mainly responsible for sickle cell crisis, is a common painful impediment of sickle cell disease in adolescents and adults. This is initiated and continued by interactions among sickle cells, endothelial cells and plasma constituents [6], vaso-occlusion is accountable for a wide variety of clinical impediments of sickle cell disease, including infraction, pain, stroke, leg ulcers,splenic infraction,priapism spontaneous abortion and renal inadequacy.

In this study we used the various biochemical factors of sickle cell crisis patients (Crisis) and compare them to that of the sickle cell disease patients in steady state (SS). The present prediction problem is a binary classification task, which is tackled using Partial least square discriminant analyis (PLSDA) and further by using random forest (RF) algorithm.

Materials and methods:

The dataset consisted of a total of 40 samples (20 belonging to SS and 20 belonging to Crisis groups). The samples are described by 20 features/descriptors (Hb-hemoglobin, HCT-Hematocrit, MCV- mean corpuscular volume, MCH- mean corpuscular hemoglobin, MCHC- mean corpuscular hemoglobin concentration, RDW-red cell distribution width, HbA- adult hemoglobin,HbA2-, HbF- fetal hemoglobin, HbS- sickle cell hemoglobin, TLC-total leukocyte count, Tbil=total bilirubin, dbill-direct billirubin). The current work involves the use of disciminatory machine learning techniques for developing predictive models for classifying SS and Crisis groups using various biochemical parameters.

Machine learning algorithms

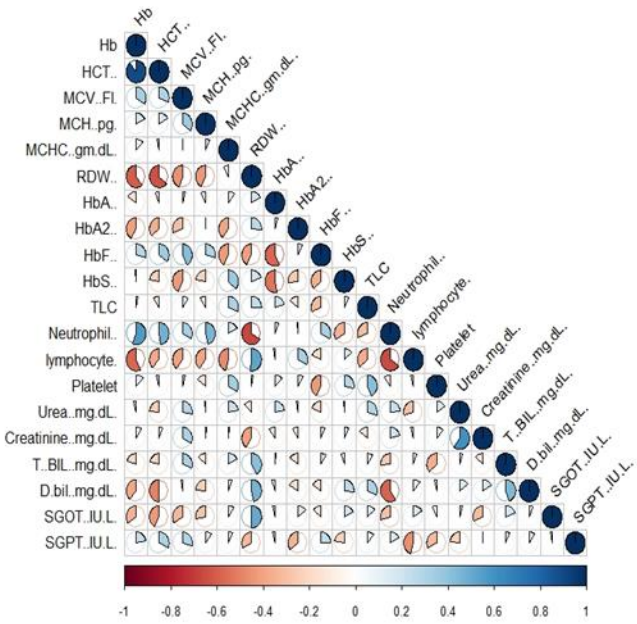
We used both partial least square discriminant analyis (PLSDA)[7-10] and further random forest (RF) algorithm for developing the discriminatory prediction models for classifying SS and Crisis groups. PLSDA is a multistep procedure which involves two majors steps: construction of PLS components and construction of discriminatory prediction model. We developed the PLSDA model with maximum of three components. Random forest [11-13] is an ensemble learning approach which conssits of traning a large number of decision tress (base classifiers) using a bootstrapped sample and random selection of features at each node of the tree. The final step completes with the fusion of individual decisions from the trained trees resulting in a final decision outcome. We used 100 decsion trees as the base learners in the random forest.

Results and discussion

The correlation plots for the different variables for the two groups are shown in Fig.1 (A and B). From Fig. 1 it can be observed that there exists different correlation patterns among the SS and Crisis groups (notably for MCH.pg, MCHC.gm.dl, Neutrophils , Lymphocytes etc.).

Volume 5, Issue 3; May-June 2022; Page No 887-894

A. Correlations among the SS group



B. Correlations among the Crisis group

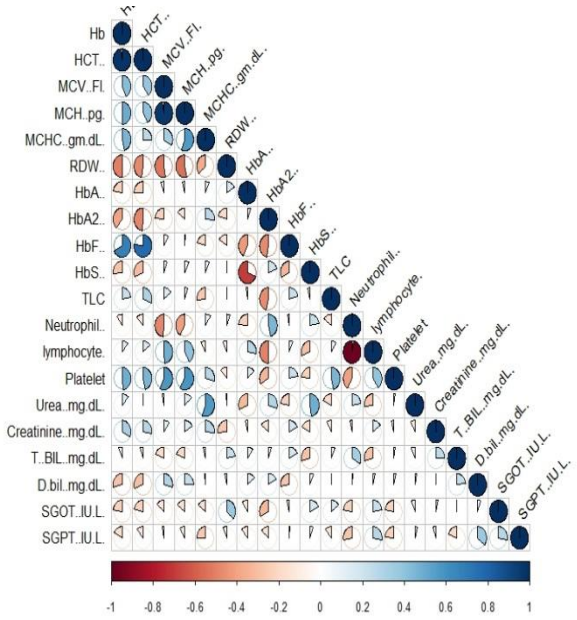


Fig.1 Showing the correlations existing among the different attributes/features present in the dataset.

The group wise distribution of the various features/attributes for the two groups is shown in Fig.2

Volume 5, Issue 3; May-June 2022; Page No 887-894

Fig.2 GroupWise boxplots for the different features/attributes of the dataset

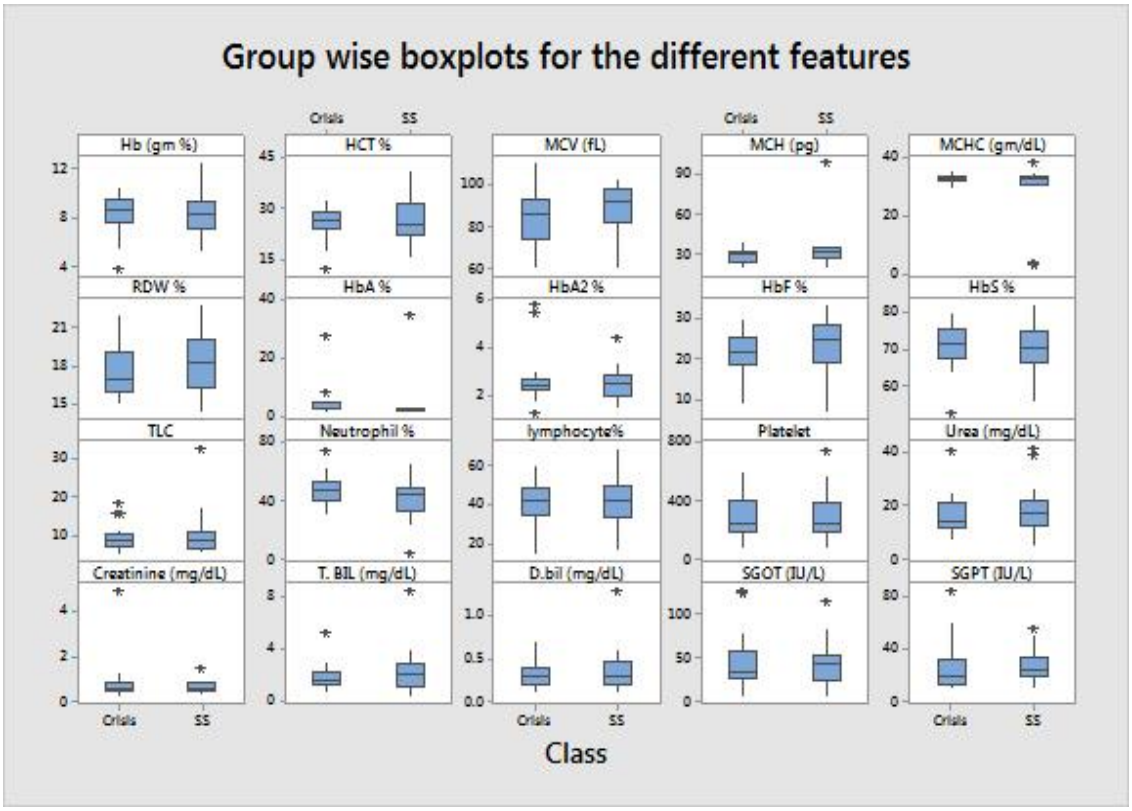


Fig.3 Correlation plot for the components, dependent and independent variables

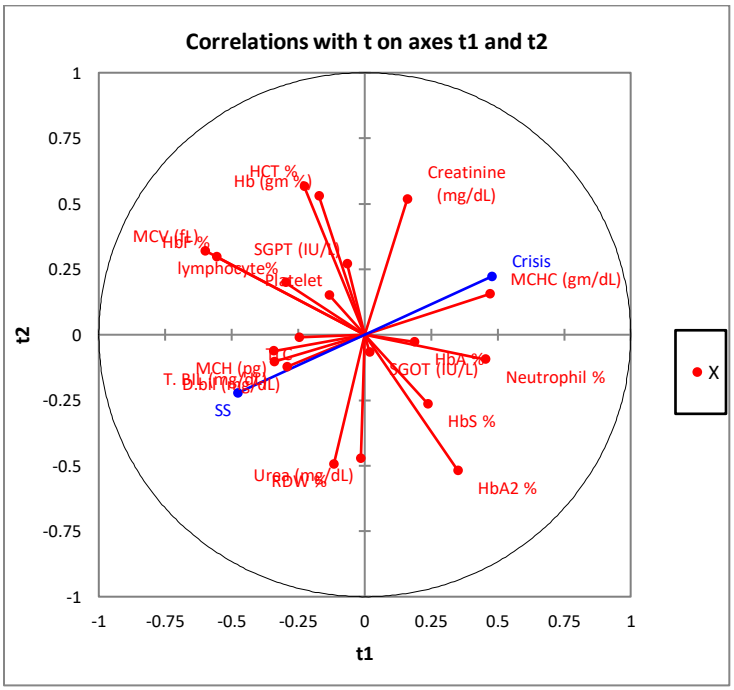


Fig.3 presents the correlation plots for the PLSDA components along with dependent and independent variables. It can be observed that Creatanine, MCHC.gm.dl and Neutrophils are more correlated with the Crisis group.

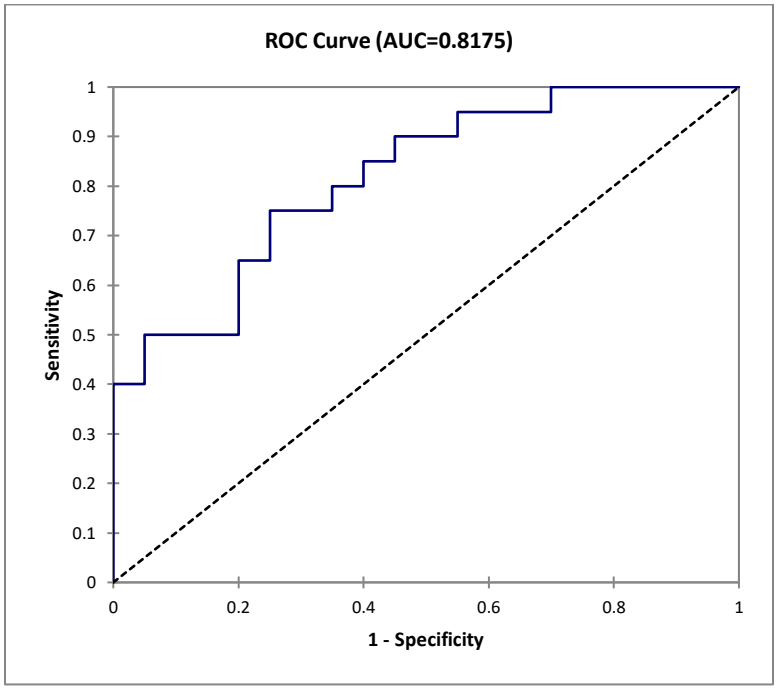
Volume 5, Issue 3; May-June 2022; Page No 887-894

The confusion matrix for the PLSDA model is presented in table 1. The PLSDA model achieved an overall accuracy of 72.5% with sensitivity of 80% and specificity of 65%. The corresponding ROC plot for the PLSDA model is shown in Fig.2. The PLSDA model obtained an AUC of value of 0.817 (fig.4).

Table 1. Confusion matrix for the PLSDA model

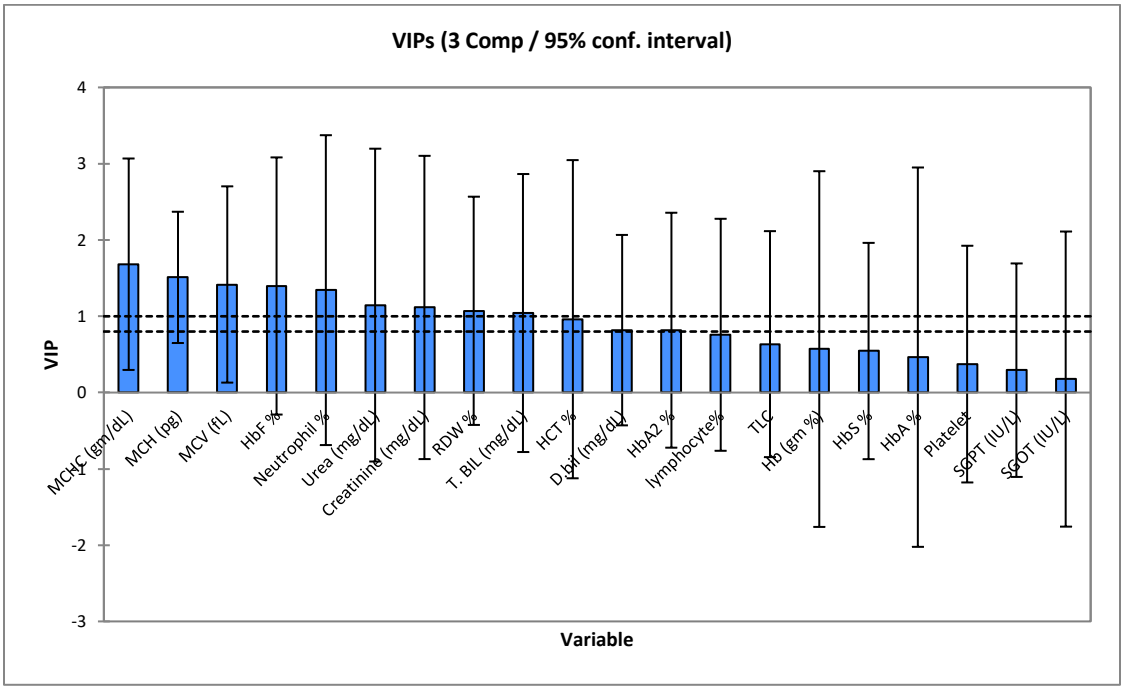
from \ to	Crisis	SS	Total	% correct
Crisis	16	4	20	80.00%
SS	7	13	20	65.00%
Total	23	17	40	72.50%

Fig.4 ROC for the PLSDA model



Volume 5, Issue 3; May-June 2022; Page No 887-894

Fig.5 Variable Importance Plots for the three components of PLSDA model



Observing the Variable importance plot (VIP) (Fig.5) for the PLSDA model with three components, it can be concluded that MCHC(gm/dl), MCH(pg), MCV, HbF % and Neutrophil % are the five most important features/attributes for the discriminatory model.

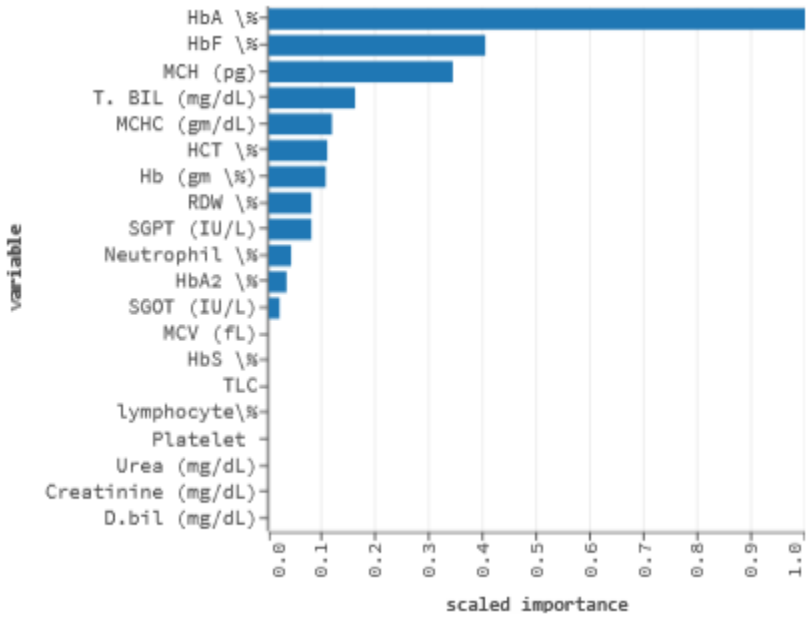
On using RF for discriminatory model generation, we obtained an overall accuracy of 82.5% with sensitivity of 70 % and specificity of 95 % , the confusion matrix is tabulated in table 2. The VIP plot for the RF discriminatory model is shown in Fig.6. For the RF model HbA, HbF, MCH (pg), T.Bil (mg/dl) and MCHC (gm/dl) are the five most important features/attributes.

Table 2 Confusion matrix for the RF model

from \ to	Crisis	SS	Total	% correct
Crisis	14	6	20	70%
SS	1	19	20	95%
Total	15	25	40	82.5%

Volume 5, Issue 3; May-June 2022; Page No 887-894

Fig. 6 Variable importance plot for the RF model



The features/attributes HbF, MCHC and MCH are the three most important features which are strongly associated with both the prediction models (PLSDA and RF).

Conclusion:

Sickle cell crisis is one of the most dreadful complications of the disease amounting to a significant burden of morbidity and mortality of the affected individuals. Incomplete understanding of the cellular mechanisms of the process has been the primary limitation for the medical fraternity in combating the sickle cell crisis. Deficiency of a reliable biomarker to diagnose and/or prognose a crisis episode leaves the physicians with the only option of clinically detecting the disorder which makes the situation even worse. Present work incorporates the implementation of machine learning algorithms for developing a discriminatory system for Steady state and Crisis patient groups using a number of biochemical parameters.

The features: HbF, MCHC and MCH are the three most important variables for the classification/discriminatory models and are strongly associated with the discrimination between the two group of patients. The current work establishes the association of biochemical parameters with severity of sickle cell condition, achieving an acceptable overall accuracy.

In future work we will implement higher order feature vectors which can be obtained using deep learning algorithms and use of feature selection methods that may facilitate in further enhancing the performance evaluation metrics of machine learning classifiers.

References

1. Rees, D.C., Williams, T.N., Gladwin, M.T., 2010. Sickle-cell disease. Lancet 376, 2018–2031.
2. Ballas, S.K., 2002. Sickle cell anemia: progress in pathogenesis and treatment. Drugs 62, 1143–1172.
3. Samuel, R.E., Salmon, E.D., Briehl, R.W., 1990. Nucleation and growth of fibers and gel formation in sickle hemoglobin. Nature 345, 833–835.
1. 4.Turnpenny P, Ellard S. Hemoglobin and the Hemoglobinopathies Emery's ELEMENTS OF MEDICAL GENETICS 14 th edition. 2012;155-164.
4. Patra P K,Chauhan V S,Khodiar P K,Dalla Al R, Serjeant G R:Screening for the sickle cell gene in Chhattisgarh state, India: an approach to a major public health problem.J Community Genet (2011) 2:147–151.
5. Steinberg MH. Management of sickle cell disease. N Engl J Med. 1999;340:1021–30.

Volume 5, Issue 3; May-June 2022; Page No 887-894

6. Sahu R, Yadav A, Nath A. Estimation of maximum recommended therapeutic dose of anti-retroviral drugs using diversified sampling and varied descriptors. Minerva Biotechnol Biomol Res 2021;33:210-8.
7. Sikka. P, et al. Inferring Relationship of Blood Metabolic Changes and Average Daily Gain With Feed Conversion Efficiency in Murrah Heifers: Machine Learning Approach. Frontiers in Veterinary Science. 2020;7:518-528.
8. Ruiz-Perez, D., Guan, H., Madhivanan, P. et al. So you think you can PLS-DA?. BMC Bioinformatics.2020; 21(2):2-12.
9. Lee LC , Liong CY , Jemain AA . Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst. 2018 Jul 23;143(15):3526-3539.
10. Nath A, Karthikeyan S, Maximizing lipocalin prediction through balanced and diversified training set and decision fusion, Computational Biology and Chemistry ,2015, 59:101-110
11. Breiman, L, Random Forests, Machine Learning, 2001, 45:5-32.
12. Nath, A, Prediction for understanding the effectiveness of antiviral peptides, Computational Biology and Chemistry,2021,95:107588

Volume 5, Issue 3; May-June 2022; Page No 887-894