EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Enhancing Disease Diagnosis: Statistical Analysis of Haematological Parameters in Sickle Cell Patients, Integrating Predictive Analytics

Bhawna Dash[1,*], Soumyalatha Naveen[2] , Ashwinkumar UM[3]

[1,2,3]School of CSE, REVA University, Bangalore, India

Abstract

Sickle cell disease (SCD) affects 30 million people worldwide, causing a range of symptoms from mild to severe, including Vaso occlusive crises (VOC). SCD leads to damaging cycles of sickling and desickling of red blood cells due to HbS polymer formation, resulting in chronic haemolytic anaemia and tissue hypoxia. We propose using machine learning to categorize SCD patients based on haemoglobin, reticulocyte count, and LDH levels, crucial markers of hemolysis. Statistical analysis, particularly Linear Regression, demonstrates how haemoglobin depletion occurs using LDH and reticulocyte parameters.

Bilirubin and haemoglobin, two integral biomarkers in clinical biochemistry and haematology, serve distinct yet interconnected roles in human physiology. Bilirubin, a product of heme degradation, is a critical indicator of liver function and various hepatic disorders, while haemoglobin, found in red blood cells, is responsible for oxygen transport throughout the body. Understanding the statistical relationship between these biomarkers has far-reaching clinical implications, enabling improved diagnosis, prognosis, and patient care. This research paper conducts a comprehensive statistical analysis of bilirubin and haemoglobin using various regression techniques to elucidate their intricate association. The primary objective of this study is to characterize the relationship between bilirubin and haemoglobin. Through meticulous data analysis, we explore whether these biomarkers exhibit positive, negative, or no correlation. Additionally, this research develops predictive models for estimating haemoglobin levels based on bilirubin data, offering valuable tools for healthcare professionals in clinical practice.

Keywords: Haemoglobin, Sickle Cell Disease, RBC, WBC, Hemoglobin, Reticulocyte, Bilirubin, Machine Learning, Regression, Clinical approach

Received on 30 December 2023, accepted on 02 April 2024, published on 09 April 2024

doi: 10.4108/eetpht.10.5691

*Corresponding author. Email: dashbhawna2000@gmail.com

1. Introduction

Hemoglobinopathy, a prevalent monogenic disorder affecting humans, poses significant genetic and social health challenges in various regions including India, South Africa, Saudi Arabia, South America, and other South Asian and African nations. Sickle cell disease (SCD) is among the oldest recognized molecular disorders, characterized by the HbS mutation (HBB Glu6Val), while HbE disease (HBB Glu26Lys) ranks as one of the most reported hemoglobin disorders after HbS. HbE arises from a mutation replacing glutamic acid with lysine at the 26th position of the Beta chain of hemoglobin.

In India, SCD predominantly affects the population in the central belt, spanning from Odisha to Maharashtra and Gujarat, with HbS, HbE, and HbDPunjab being the three main genotypes observed. Anemia, the most common blood condition, results from a deficiency of red blood cells (RBCs), impairing the body's oxygen supply. Acute anaemia stems from a sudden drop in RBC count, while chronic anaemia develops gradually, often accompanying inflammatory conditions. SCD disrupts normal RBC

formation, leading to characteristic sickle-shaped cells and a shortened lifespan for RBCs.

Symptoms of sickle cell anaemia typically manifest around five months of age, marked by RBC destruction and reduced circulation. Common manifestations include edema, frequent infections, eye issues, stunted growth, and delayed puberty. SCA results from the abnormal haemoglobin type, HbS, inherited from both parents, leading to the sickle cell homozygote phenotype.

While SCD has no cure, management strategies aim to alleviate symptoms and prevent complications through regular clinical and hematobiochemical monitoring. Machine learning techniques offer promise in predicting and assessing the health status of SCD patients, aiding in early intervention to mitigate future complications. This study focuses on utilizing machine learning tools to analyse data from a patient database in Odisha, aiming to predict and evaluate the health status of individuals suffering from sickle cell anemia. Bilirubin and haemoglobin, although distinct in their functions and origins, are vital elements in the complex web of human physiology. Bilirubin, a yellow pigment formed during the breakdown of heme, serves as a crucial marker for evaluating liver function and diagnosing a wide spectrum of hepatic disorders, including jaundice, hepatitis, and cirrhosis [1]. In contrast, haemoglobin, an iron-containing protein found in red blood cells, plays an indispensable role in the transport of oxygen throughout the body, ensuring the oxygenation of tissues and organs [2]. Given their central roles in health and disease, a detailed statistical analysis of the relationship between bilirubin and haemoglobin becomes imperative.

This research paper embarks on an in-depth exploration of the statistical analysis of bilirubin and haemoglobin, employing a range of regression techniques as powerful analytical tools. The fundamental objective of this study is to unravel the intricate interplay between these two biomarkers. This knowledge holds immense clinical significance, offering insights into the pathophysiological mechanisms underpinning a myriad of medical conditions. Moreover, the research endeavours to construct predictive models capable of estimating haemoglobin levels based on bilirubin data. Such models hold substantial promise in clinical practice, facilitating early diagnosis, treatment monitoring, and individualized patient care.

The clinical relevance of this research paper is profound and multifaceted. By delving into the statistical nuances of bilirubin and haemoglobin, this study contributes to a deeper understanding of the pathophysiology of various diseases. Moreover, it equips healthcare professionals with robust analytical tools that can aid in early detection, prognosis, and therapeutic decision-making. The development of predictive models is not undertaken lightly; rigorous statistical validation ensures the accuracy and generalizability of these models across diverse patient populations. As the statistical intricacies of bilirubin and hemoglobin are elucidated, this research aims to empower medical practitioners with knowledge and resources to enhance patient outcomes. Ultimately, the findings of this study promise to improve the quality of care, offering new avenues for diagnosis and treatment in the realm of liver diseases, anaemia, and related medical conditions.

2. Literature Review

Sen et al. (2021) [3] conducted a study utilizing microscopic blood samples and employing techniques such as image processing and machine learning to automate the detection of sickle cells. They categorized the detected RBCs based on shape into three categories: circular, elongated sickle cell-shaped, and others. The images were pre-processed, and segmentation was performed using the Otsu thresholding technique.

Petrović et al. (2020) [4] utilized peripheral blood smears to observe RBC images, which were then segmented and classified based on morphology using machine learning techniques.

Nkpordee (2022) [5] conducted a case study in Nigeria, employing time series trend models and statistics to project SCD prevalence over six years and predict a decline in the coming years.

Patel et al. (2021) [6] emphasized the importance of early sickle cell detection for symptom identification and subsequent management. They employed data mining techniques, including classification algorithms, to accurately identify sickle cells in the human body.

Yang (2018) [7] and Yeruva (2021) [8] utilized machine learning algorithms to predict hospital readmissions in SCD cases. They partitioned patients into testing and training groups and evaluated predictions using various algorithms and metrics.

Dean (2019) [9] used Multinomial Logistic Regression to analyse pain scores in SCD patients, developing a machine learning model to predict pain scores effectively.

Wing (2019) [10] proposed a low-cost, non-invasive sickle cell screening device suitable for use in developing countries.

Stone (2021) [11] presented a case report highlighting the severity of a delayed haemolytic transfusion reaction in an SCD patient undergoing red cell exchange for gene therapy.

Ranjana (2020) [12] explored automatic categorization of SCA using image processing techniques and machine learning, achieving high classification accuracy.

Patgiri (2022) [13] demonstrated a hybrid segmentation procedure combining fuzzy C-means segmentation with adaptive thresholding for analysing blood smear samples, achieving promising results with supervised classifiers.

Machine learning algorithms have the potential to be employed in forecasting Acute Kidney Injury (AKI) [14] in individuals diagnosed with Sepsis-Associated Acute Respiratory Distress Syndrome (ARDS). As a result, an easily navigable Shiny application, utilizing the XGBoost model known for its trustworthy predictive accuracy, has been made accessible on the internet. This application aims to estimate the likelihood of AKI occurrence in patients who have been diagnosed with sepsis-associated ARDS.

This study [15] involved 329 patients with delayed Methotrexate (MTX) elimination and 1400 patients without this issue, all meeting the inclusion criteria. Using univariate and LASSO regression, eleven predictors were identified, including age, weight, creatinine, uric acid, total bilirubin, albumin, white blood cell count, haemoglobin, prothrombin time, immunological classification, and omeprazole co-medication. The XGBoost algorithm with SMOTE showed strong performance with an AUROC of 0.897, AUPR of 0.729, sensitivity of 0.808, and specificity of 0.847, outperforming other models. External validation confirmed an AUROC of 0.788.

The authors in this study [16] conveyed how acute heart failure (AHF) is a common and severe condition often complicated by worsening renal function (WRF), worsening the prognosis. They have used clustering, a machine learning technique, on data from 312 AHF patients with 86 variables and identified three distinct patient clusters with significantly different WRF incidences ($p = 0.004$).

In this research paper, the authors [17] have demonstrated the efficiency of a non-invasive technique for the detection of jaundice, which offers a more comfortable experience compared to conventional methods. The primary objective of the authors is to deliver bilirubin test results and patient treatment in a cost-effective and expedited manner, addressing the specific needs of such healthcare settings.

This study addresses [18] blood sample rejection due to haemolysis in clinical labs, developing a cost-effective method for haemolysis detection in small plasma volumes, even with bilirubin and lipid interferents. Experimental samples used plasma from whole blood with varying haemoglobin, bilirubin, and lipid concentrations. An optical setup measured haemoglobin from 0 to 400 mg/dL using <1 µL of detection volume, achieving >90% sensitivity with ~10% variation across 27 samples. This approach's sub-microliter detection and high sensitivity hold potential for point-of-care medical devices, mitigating haemolysis-related inaccuracies in clinical measurements.

The authors [19] used modern machine learning to optimize sensor performance, introducing a behaviour-predicting sensor through polynomial regression. They explored three meta surface variations, aiming for maximum sensitivity. Notably, the double split-ring resonator and single split-ring resonator designs showed the highest sensitivity. The study also analysed parameter changes, like thickness, affecting absorption. A polynomial regression model efficiently predicted absorption values, with consistently high R2 scores above 5, demonstrating its accuracy. This biosensor, designed using the PR model, holds promise for biomedical applications, including haemoglobin detection.

In this study [20], the methodology comprises three phases: dataset collection of palm images, image preprocessing (including extraction, augmentation, ROI segmentation, and color space analysis), and the development of anemia detection models using various algorithms (CNN, k-NN, Naive Bayes, SVM, and Decision Tree). The initial dataset had 527 samples, expanded to 2635 through augmentation (rotation, flipping, and translation). This augmented dataset was randomly divided into training (70%), validation (10%), and testing (20%) sets for model assessment.

Within this research [21], the diagnosis of various forms of anaemia is identified as a resource-intensive endeavour, often demanding subsequent costly examinations due to the inherent constraints of the complete blood count (CBC). Smaller healthcare establishments may encounter challenges in accessing specialized diagnostic tools, thereby complicating the differentiation of anaemia types, including beta thalassemia trait (BTT), iron deficiency anaemia (IDA), haemoglobin E (HbE), and combination anaemias. To address this intricate issue, the authors have developed a meticulously precise automated prediction model through the utilization of the extreme learning machine (ELM) algorithm. Remarkably, this model has demonstrated noteworthy achievements, boasting a remarkable 99.21% accuracy, 98.44% sensitivity, 99.30% precision, and an F1 score of 98.84%, all of which were attained by harnessing historical data.

In this study [22], 158 cases of end-stage renal disease (ESRD) occurred, constituting 40.51% of cases over a 3-year follow-up. The Random Forest (RF) algorithm excelled in predicting ESRD progression, achieving an impressive AUC of 0.90 and an accuracy rate of 82.65%. The RF algorithm identified five key predictors: Cystatin-C, serum albumin (sAlb), haemoglobin (Hb), 24-hour urine urinary total protein, and estimated glomerular filtration rate. A predictive nomogram was developed using these factors to anticipate ESRD incidence.

In their study [23], the authors identified and analyzed 2935 patients from the MIMIC-III database and an additional 499 patients from their local database to create and validate the AKI risk prediction model. The incidence of AKI in these two cohorts was 18.3% and 61.7%, respectively. Key factors associated with AKI included

various laboratory parameters, age, and hospital stay duration. The analysis revealed that the XGBoost model performed the best, with an AUROC of 0.880 (95%CI: 0.831–0.929), indicating its superior ability to predict AKI risk in patients with acute cerebrovascular disease.

The authors [24] detailed that both systemic and non-systemic conditions leading to iron deposition in the spleen exhibit distinct MRI characteristics due to the susceptibility effects of deposited iron or hemosiderin. Dual-echo gradient-echo sequences reveal signal loss on longer echo sequences attributable to the T2* effect. Understanding the pathophysiology of these iron sequestration diseases, whether systemic or localized, can aid in diagnosing underlying conditions when combined with clinical data.

In this study, the authors [25] investigated human erythrocytes, essential for oxygen transport. These organelle-free cells, numbering around 25 trillion per person, have a 120-day lifespan, contributing significantly to tissue homeostasis. While erythropoiesis mechanisms are well-understood, erythrocyte clearance, particularly eryptosis, was the focus. Eryptosis involves factors like Ca2+ influx, ceramide generation, oxidative stress, kinase activation, and iron metabolism. The study also explored parallels with ferroptosis, an iron-dependent death of nucleated blood cells, and its relevance in infectious diseases and hematologic disorders.

In this study [26], the authors created a model to predict phototherapy need in 98 out of 362 neonates. Achieving an impressive 95.20% accuracy, the model forecasts treatment necessity up to 48 hours in advance using just four key variables: bilirubin levels, weight, gestational age, and hours since birth. This tool, named the early phototherapy prediction tool (EPPT), is available as an open web application.

3. Linear Regression

Linear Regression is a well-known supervised learning algorithm used for predicting dependent variables based on given independent variables. When given a set of independent variables, Logistic Regression is employed for categorical prediction of dependent variables. It can be represented as:

$$y = c0 + c1x+e \quad (1)$$

Here, y is the dependent variable, x is the independent variable, c0 is the constant term and intercept, c1 is the regression coefficient or slope, and e represents the random error, as illustrated in Figure 1.

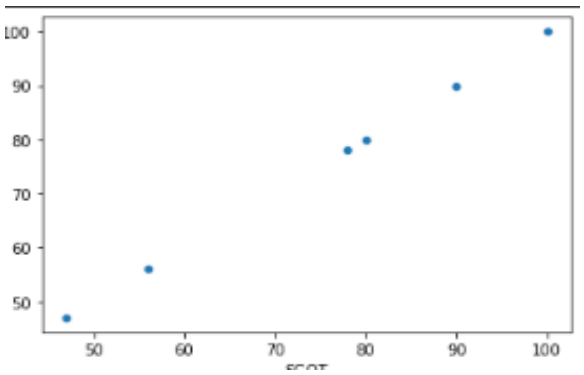

Fig. 1. Linear regression graph

The goal is to determine the best values for c0 and c1 to minimize the error between predicted and actual values, expressed in equations 2 and 3:

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

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

This function aims to minimize the squared error between actual and predicted values, commonly known as Mean Squared Error (MSE). Q represents the average squared error across all data points. The values of c0 and c1 are adjusted iteratively to minimize MSE.

Figure 2 illustrates the workflow of fitting a linear regression model. It involves importing data, employing an optimization technique and cost function for performance evaluation, adjusting parameters to improve quality, and obtaining output.

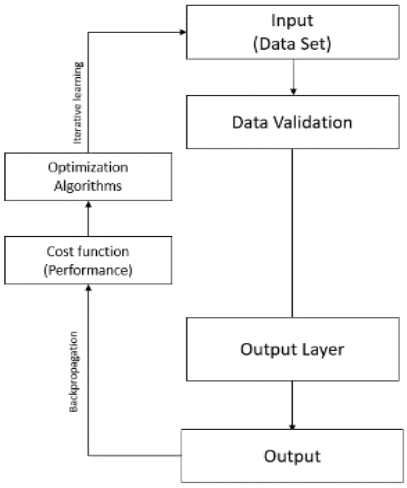

Fig. 2. Flowchart of the working of linear regression while using a dataset.

Stochastic Gradient Descent (SGD) is a variant of gradient descent used to optimize machine learning models. It calculates the gradient and updates parameters

using only one random training example per iteration. This algorithm is beneficial when optimal points cannot be found by setting the slope of the function to zero. In linear regression, the sum of squared residuals is mentally mapped as the function "y," and the weight vector represents "x" in the parabola.

4. Proposed Work

In the proposed approach we have used sickle cell patient dataset from a hospital in Western Odisha. The dataset comprises 100 rows and 5 columns, which have been extracted following preprocessing steps. These columns encompass a range of essential parameters including White Blood Cell (WBC) count, Red Blood Cell (RBC) count, Haemoglobin (HGB) levels, as well as parameters denoted as BIT, BID, LDH, and RETICS%. These parameters collectively represent key indicators of health and disease progression in individuals affected by sickle cell disease. Through meticulous preprocessing, the data has been refined to ensure accuracy and reliability, laying the groundwork for comprehensive analysis and insights into the physiological profiles of these patients. The extracted data was analysed and used as an input for our machine learning algorithm models. Various regression models were used, and the best of the model will be used further for analysis. So, we transformed and cleaned the data before feeding it to the suitable algorithms for both testing and training as shown in Fig.3. Following the visualisation and train-test split, we choose an appropriate model to perform our intended task. We arrived at an appropriate conclusion after acquiring the accuracy and correct graphs (as shown in the Graphs and Results section).

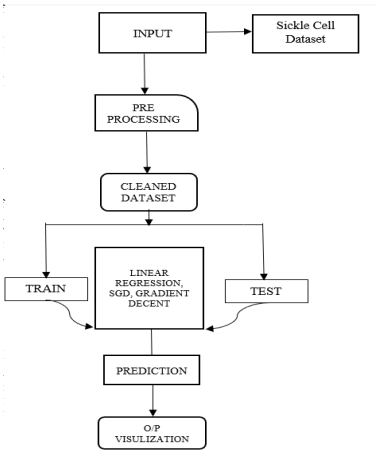

Fig. 3. Flowchart describing the methodology that was used in this study.

Fig 4 illustrates the deployment process of the regression models employed in the paper, which include Linear Regression, Decision Tree Regression, and Weighted Averaging Models.

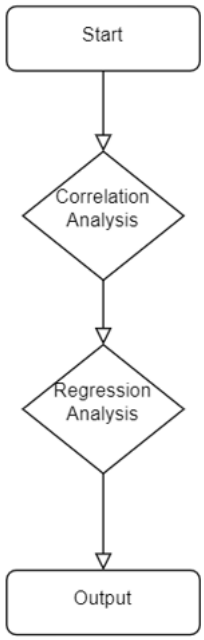

Fig 4: Flowchart depicting the statistical analysis used for deploying various regression models.

Leveraging statistical methodologies provides a robust framework for the classification, prediction, and development of optimal models aimed at assessing the health and clinical status of individuals afflicted with sickle cell anaemia or experiencing reduced haemoglobin levels. These statistical tools enable in-depth analyses of patient data, encompassing various factors such as genetic profiles, blood cell characteristics, and clinical manifestations. Through the application of techniques like linear regression, logistic regression, and gradient descent, researchers can construct predictive models capable of evaluating disease severity, progression, and potential complications.

Moreover, statistical methods offer a means to identify biomarkers and risk factors associated with these conditions, facilitating early detection and intervention efforts. By employing meticulous data preprocessing, feature selection, and model validation procedures, the accuracy and reliability of these predictive models are bolstered. This holistic approach empowers healthcare practitioners to tailor personalized treatment strategies, monitor disease trajectories, and optimize therapeutic interventions for individuals grappling with hemoglobinopathies or related ailments.

By expanding the integration of statistical methodologies within healthcare contexts, we can unlock deeper insights into the intricate nature of these disorders, thereby fostering more effective clinical management and fostering improved patient outcomes.

The pseudocode for the methodology proceeds as shown in Table 1 below.

TABLE 1. Pseudocode of proposed work

Algorithm: Predicting Health/Clinical Status of Individuals with Sickle Cell Anaemia or Low Haemoglobin

Input: Patient dataset

Data preprocessing

Process:

1. Divide the dataset into Train (80%) and Test (20%) sets.

2. For each data in the dataset:

a. Apply Linear Regression on Train and Test data.

b. Implement gradient descent.

c. Predict the test results.

Output: Predicted percentages, Correlation matrix

5. Results

The proposed work under section IV has been analysed using various regression models and the difference between the results were hence derived. The dataset has been collected from the western part of Odisha which is highly affected by sickle cell disease patients.

Data was gathered from the western region of Odisha state, a region highly affected by Sickle Cell Disease (SCD). The dataset comprises six distinct categories: White Blood Cell count (WBC), Red Blood Cell count (RBC), Hemoglobin concentration (HGB), Bilirubin Induced Transcription factor (BIT), Lactate Dehydrogenase (LDH), and Reticulocyte Percentage (RETICS%). All values, except for RETICS%, have been normalized to per unit levels.

To mitigate redundancy, the dataset underwent analysis with a novel value set at 0.17. Scatter plot analyses were conducted using LDH and RETICS% as target variables, with the other four parameters serving as predictors. The effects of LDH and RETICS% were further analyzed in relation to HGB levels.

Of the total dataset, 80% was allocated for model training, while the remaining 20% was reserved for testing purposes.

Table 2 presents the statistical results from three distinct machine learning (ML) algorithms: Decision Tree, Linear Regression, and Support Vector Machine. Notably, Linear Regression exhibited the lowest Root Mean Squared Error (RMSE) of 3.60, with an R-squared error of -0.39. However, the Mean Absolute Error (MAE) was calculated at 2.72, potentially influenced by the homogeneity of data within the training dataset. Subsequent analysis was performed using Linear Regression with Gradient Descent.

TABLE. 2. 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

In Figure 5, the regression analysis depicts the relationship between RETICS and HGB. The majority of the data points exhibit characteristics of a negative slope, indicating that as the HGB content increases, there is a corresponding decrease in RETICS levels. Across most cases, the HGB content falls within the range of 8.5 to 11, corresponding to a decrease of approximately 17% in RETICS.

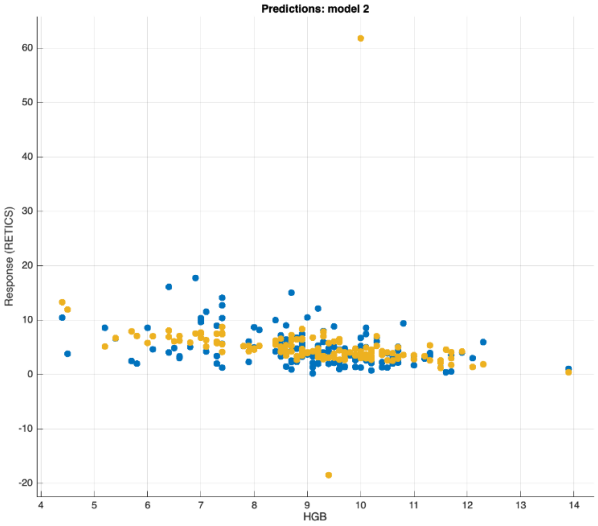

Fig. 5. Regression analysis of RETICS vs HGB

Similarly, Figure 6 illustrates the statistical graphical analysis depicting the relationship between LDH and HGB. Notably, within the same range of HGB values as seen in Figure 5, which ranges from 8.5 to 11, the LDH content fluctuates between 385 to 460.

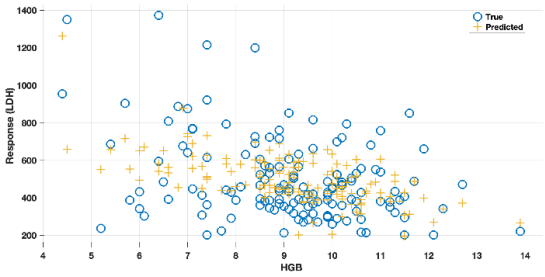

Fig. 7. Regression analysis of LDH vs HGB

Table 3 presents the statistical analysis results of various machine learning (ML) algorithms applied to the LDH vs HGB relationship. Meanwhile, Figure 8 displays a heatmap generated based on the autocorrelation function, where each diagonal element possesses a magnitude of 1 per unit.

TABLE. 3. Statistical Analysis of ML Algorithms

	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

This correlation suggests as shown in Table 4 that individuals who exhibit higher levels of LDH relative to HGB may indicate insufficient hemoglobin production in their bodies. Consequently, this imbalance may impede their capacity to adequately transport oxygen throughout their system.

TABLE. 4. 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

One of the important parameters that was analysed here was BIT(Bilirubin) with conjunction with Haemoglobin (Hb). The goal of this study was to use several regression approaches to investigate the relationship between bilirubin and haemoglobin levels. According to the findings of this study, there was a very weak negative correlation between bilirubin and haemoglobin levels, as assessed by regression analysis. Furthermore, the data imply that this association may vary between people, as evidenced by increase in bilirubin concentrations in the gallbladder.

Redundancy analysis was conducted on the processed data, resulting in an assigned new value of 0.17. Subsequently, a scatter plot analysis was executed on a target data, BIT, leveraging four additional parameters as predictors. HGB was employed to stratify the influence of BIT. The dataset was partitioned, allocating 80% for training purposes and the remaining 20% for testing the model.

Table 5 shows the statistical analysis of various regression analysis such as Decision Tree Regression, Linear Regression and Weighted averaging Regression models.

Table.6:- Various regression models evaluated based on MSE

Models	MSE
DTR	8.737012440376152
LR	6.669681797743141
WAR	7.739959209300224

It was found that all the models showed a very weak negative correlation among which DTR (Fig 8) and LR (Fig 9) showed a very weak negative corelation existed between the two parameters, and the WAR visually shows no proper correlation, visually (Fig 10).

The results show that with the decreasing amount of bilirubin in patients' body, haemoglobin levels are increasing.

Hence this will prove that due to constant haemolysis in a sickle cell patients' body bilirubin level falls which leads to jaundice and various other renal problem in a SCD patient. Hence as compared to a normal patient the bilirubin level in a SCD patient is very high i.e. below 300mg/dL but compared to normal patients' whole bilirubin levels lie between 0.5mg/dL to 1mg/dL.

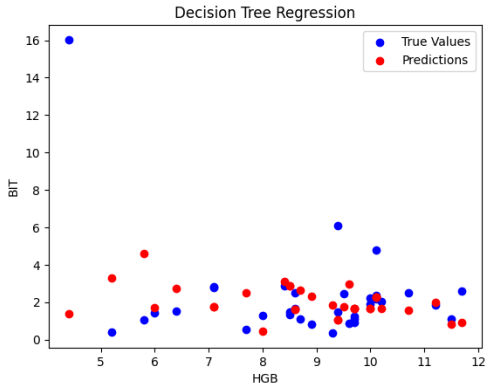

Fig 8. Curve Fitting graph for decision Tree Regression analysis

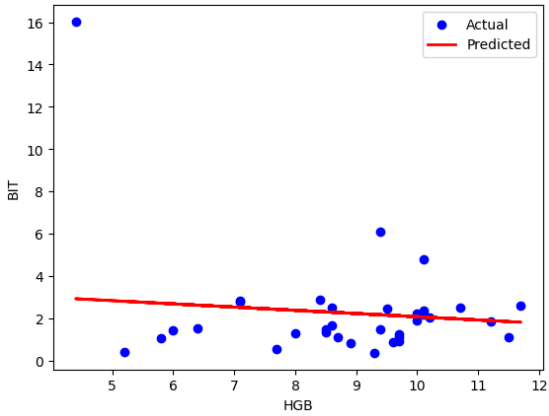

Fig 9. Curve Fitting Graph for Linear Regression analysis

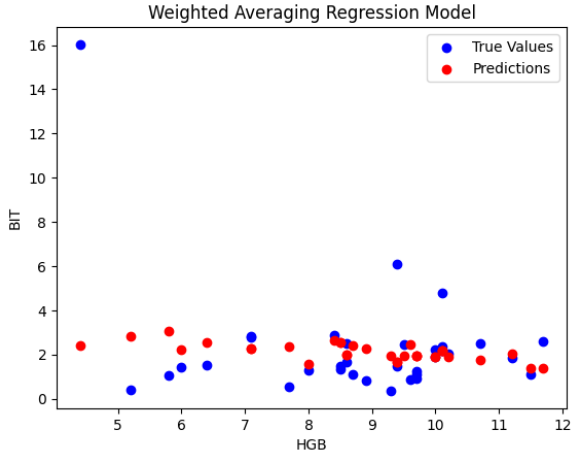

Fig 10. Curve Fitting Graph for Weighted Averaging Regression Analysis

6. Conclusion and Future Work

In conclusion, our study, utilizing machine learning models, predicts that patients with sickle cell disorder (HbSS) and low hemoglobin levels may experience various clinical symptoms attributed to elevated LDH levels, increased WBC counts, and reticulocyte counts. Their reduced hemoglobin levels hinder oxygen transport, leading to a deoxygenated state, elevated lactic acid, and heightened reticulocyte counts. Consequently, regular blood transfusions and hemoglobin tests are recommended for effective management.

Furthermore, introducing non-invasive methods for hemoglobin measurement is crucial to alleviate patient discomfort associated with traditional painful methods. Meanwhile, leveraging machine learning techniques for managing patient data is essential for effective monitoring and treatment.

Moreover, our comprehensive statistical analysis of bilirubin and hemoglobin relationships provides valuable insights, establishing correlations and predictive models with clinical significance. Future research endeavors should aim to expand upon these findings, incorporating additional biomarkers, conducting rigorous clinical validations, and exploring longitudinal studies to deepen our understanding of disease dynamics and treatment responses.

The application of advanced machine learning techniques holds promise for enhancing predictive capabilities, contributing to improved diagnostics and patient care across various medical conditions. As we continue to explore the complexities of bilirubin and hemoglobin, our efforts aim to advance medical knowledge and enhance patient outcomes.

Acknowledgements.

I would like to thank Dr. Bisnu Prasad Dash, Dr. Sasmita Dash, Dr. Satyabrata Meher and Mr. A Sai Rajesh for contributing their time and providing the dataset, as well as for supervising and understanding sickle cell disease based on their study findings. Moreover I would like to thank REVA University for giving me the opportunity to proceed with the research.

References

[1] V. Narwal et al., “Bilirubin detection by different methods with special emphasis on biosensing: A review,” Sensing and Bio-Sensing Research, vol. 33. Elsevier BV, p. 100436, Aug. 2021. doi: 10.1016/j.sbsr.2021.100436.

[2] A. Atipimonpat et al., “Extracellular vesicles from thalassemia patients carry iron-containing ferritin and hemichrome that promote cardiac cell proliferation,” Annals of Hematology, vol. 100, no. 8. Springer Science and Business Media LLC, pp. 1929–1946, Jun. 21, 2021. doi: 10.1007/s00277-021-04567-z

[3] B. Sen, A. Ganesh, A. Bhan, S. Dixit and A. Goyal, "Machine learning based Diagnosis and Classification Of Sickle Cell Anemia in Human RBC," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 753-758, doi: 10.1109/ICICV50876.2021.9388610.

[4] 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, http://dx.doi.org/10.1016/j.compbiomed.2020.104027,

(2020). https://doi.org/10.1016/j.compbiomed.2020.104027.

[5] Nkpordee, L., & Wonu, N. (2022). Statistical modelling of genetic disorder in Nigeria: a study of sickle cell disease. *Faculty of Natural and Applied Sciences Journal of Scientific Innovations*, *3*(2), 10–19. Retrieved from https://www.fnasjournals.com/index.php/FNAS-JSI/article/view/27

[6] Patel, A., Gan, K., Li, A. A., Weiss, J., Nouraie, M., Tayur, S., & Novelli, E. M. (2020). Machine-learning algorithms for predicting hospital re-admissions in sickle cell disease. In British Journal of Haematology (Vol. 192, Issue 1, pp. 158–170). Wiley. https://doi.org/10.1111/bjh.17107.

[7] Yang, F., Banerjee, T., Narine, K., & Shah, N. (2018). 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. https://doi.org/10.1016/j.smhl.2018.01.002

[8] Yeruva, S., Gowtham, B. P., Chandana, Y. H., Varalakshmi, M. S., & Jain, S. (2021). Prediction of Anemia Disease Using Classification Methods. In Machine Learning Technologies and Applications (pp. 1–11). Springer Singapore. https://doi.org/10.1007/978-981-33-4046-6_1

[9] Dean, C. L., Maier, C. L., Chonat, S., Chang, A., Carden, M. A., El Rassi, F., McLemore, M. L., Stowell, S. R., & Fasano, R. M. (2019). Challenges in the treatment and prevention of delayed hemolytic transfusion reactions with hyperhemolysis in sickle cell disease patients. In Transfusion (Vol. 59, Issue 5, pp. 1698–1705). Wiley. https://doi.org/10.1111/trf.15227

[10] J. Wing et al., "A Low-Cost, Point-of-Care Sickle Cell Anemia Screening Device for Use in Low and Middle-Income Countries," 2019 IEEE Global Humanitarian Technology Conference (GHTC), Seattle, WA, USA, 2019, pp. 1-4, doi: 10.1109/GHTC46095.2019.9033017.

[11] Stone, E.F., Avecilla, S.T., Wuest, D.L., Lomas-Francis, C., Westhoff, C.M., Diuguid, D.L., Sadelain, M., Boulad, F., Shi, P.A.: 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, http://dx.doi.org/10.3324/haematol.2020.253229, (2020). https://doi.org/10.3324/haematol.2020.253229

[12] Ranjana, S., R. Manimegala, and K. Priya.: Automatic Classification of Sickle Cell Anemia using Random Forest Classifier. In: Proceedings of the European Conference on Medical Advances, LNCS, vol. 9999, pp. 2020. Springer, Heidelberg (2020).

[13] 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,http://dx.doi.org/10.1016/j.bspc.2021.102745,(2021).https://doi.org/10.1016/j.bspc.2021.102745.

[14] Y. Zhou et al., "MACHINE LEARNING MODELS FOR PREDICTING ACUTE KIDNEY INJURY IN PATIENTS WITH SEPSIS-ASSOCIATED ACUTE RESPIRATORY DISTRESS SYNDROME," Shock, vol. 59, no. 3. Ovid Technologies (Wolters Kluwer Health), pp. 352–359, Jan. 10, 2023. doi: 10.1097/shk.0000000000002065.

[15] C. Jian et al., "Predicting delayed methotrexate elimination in pediatric acute lymphoblastic leukemia patients: an innovative web-based machine learning tool developed through a multicenter, retrospective analysis," BMC Medical Informatics and Decision Making, vol. 23, no. 1. Springer Science and Business Media LLC, Aug. 03, 2023. doi: 10.1186/s12911-023-02248-7.

[16] S. Urban et al., "Machine Learning Approach to Understand Worsening Renal Function in Acute Heart Failure," Biomolecules, vol. 12, no. 11. MDPI AG, p. 1616, Nov. 02, 2022. doi: 10.3390/biom12111616.

[17] Md. M. M. Miah et al., "Non-Invasive Bilirubin Level Quantification and Jaundice Detection by Sclera Image Processing," 2019 IEEE Global Humanitarian Technology Conference (GHTC). IEEE, Oct. 2019. doi: 10.1109/ghtc46095.2019.9033059.

[18] M. Azhar et al., "Hemolysis Detection in Sub-Microliter Volumes of Blood Plasma," in IEEE Transactions on Biomedical Engineering, vol. 67, no. 5, pp. 1243-1252, May 2020, doi: 10.1109/TBME.2019.2934517.

[19] S. K. Patel, J. Surve, J. Parmar, A. Natesan and V. Katkar, "Graphene-Based Metasurface Refractive Index Biosensor for Hemoglobin Detection: Machine Learning Assisted Optimization," in *IEEE Transactions on NanoBioscience*, vol. 22, no. 2, pp. 430-437, April 2023, doi: 10.1109/TNB.2022.3201237.

[20] P. Appiahene, J. W. Asare, E. T. Donkoh, G. Dimauro, and R. Maglietta, "Detection of iron deficiency anemia by medical images: a comparative study of machine learning algorithms," BioData Mining, vol. 16, no. 1. Springer Science and Business Media LLC, Jan. 24, 2023. doi: 10.1186/s13040-023-00319-z.

[21] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, "A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia," Healthcare, vol. 11, no. 5. MDPI AG, p. 697, Feb. 26, 2023. doi: 10.3390/healthcare11050697

[22] Y. Zou et al., "Development and internal validation of machine learning algorithms for end-stage renal disease risk prediction model of people with type 2 diabetes mellitus and diabetic kidney disease," Renal Failure, vol. 44, no. 1. Informa UK Limited, pp. 562–570, Apr. 04, 2022. doi: 10.1080/0886022x.2022.2056053.

[23] X. Zhang, S. Chen, K. Lai, Z. Chen, J. Wan, and Y. Xu, "Machine learning for the prediction of acute kidney injury in critical care patients with acute cerebrovascular disease," Renal Failure, vol. 44, no. 1. Informa UK Limited, pp. 43–53, Feb. 15, 2022. doi: 10.1080/0886022x.2022.2036619.

[24] N. Consul, S. Javed-Tayyab, A. C. Morani, C. O. Menias, M. G. Lubner, and K. M. Elsayes, "Iron-containing pathologies of the spleen: magnetic resonance imaging features with pathologic correlation," Abdominal Radiology, vol. 46, no. 3. Springer Science and Business Media LLC, pp. 1016–1026, Sep. 11, 2020. doi: 10.1007/s00261-020-02709-x.

[25] P. Dreischer, M. Duszenko, J. Stein, and T. Wieder, "Eryptosis: Programmed Death of Nucleus-Free, Iron-Filled Blood Cells," Cells, vol. 11, no. 3. MDPI AG, p. 503, Feb. 01, 2022. doi: 10.3390/cells11030503.

[26] B. D. Kamath-Rayne, E. A. DeFranco, and M. P. Marcotte, "Antenatal Steroids for Treatment of Fetal Lung Immaturity After 34 Weeks of Gestation," Obstetrics & Gynecology, vol. 119, no. 5. Ovid Technologies (Wolters Kluwer Health), pp. 909–916, May 2012. doi: 10.1097/aog.0b013e31824ea4b2. Available: http://dx.doi.org/10.1097/AOG.0b013e31824ea4b2