Journal of Information Systems and Informatics
Vol. 5, No. 4, December 2023 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v5i4.629 Published By DRPM-UBD

Machine Learning Approach for Classification of Sickle Cell Anemia in Teenagers Based on Bayesian Network

Blessing Ekong[1], Otuekong Ekong[2], Abasiama Silas[3], Anthony Effiong Edet[4,*], Bright William[5]

[1,2,3,4,5]Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria
Email: [1]blessingekong@aksu.edu.ng, [2]otuekongekong@aksu.edu.ng, [3]abasiamasilas@aksu.edu.ng, [4]anthonyedet73@gmail.com

Abstract

In this study, we employed a Bayesian network approach for the classification of sickle cell anemia in teenagers based on their medical data. Sickle cell anemia is a hereditary blood disorder characterized by the presence of abnormal hemoglobin, leading to distorted red blood cells. Early detection and classification of this condition are crucial for timely intervention and improved patient outcomes. Our research focused on leveraging the algorithmic power of Bayesian network to model and analyze a diverse set of medical parameters in teenagers, including age, platelet count, mean corpuscular hemoglobin concentration (MCHC), red blood cell count, packed cell volume etc. The Bayesian network method employed for the classification of sickle cell anemia involves using a probabilistic graphical model to represent the relationships among different medical parameters. The model incorporates Bayesian principles to update and refine its predictions as new information is introduced. The method identifies key features in the dataset that contribute significantly to the classification, providing valuable insights for early detection and intervention. The Bayesian network demonstrated remarkable efficacy in accurately classifying teenagers as either positive for sickle cell anemia or negative, achieving an impressive 99% accuracy rate. This high level of accuracy indicates the robustness of the model in discerning intricate patterns within the medical data. Key features contributing to the classification are found in the dataset, shedding light on their relevance in distinguishing between positive and negative cases of sickle cell anemia, especially in teenagers. Our findings provide valuable insights into the potential diagnostic significance of sickle cell anemia classification in the teenage population. This research contributes to the growing body of knowledge in the field of medical informatics and computational biology, offering an efficient and reliable tool for healthcare practitioners in the early identification of sickle cell anemia in teenagers. The demonstrated accuracy of the Bayesian network shows its potential as an effective decision support system, aiding clinicians in making informed decisions and facilitating timely interventions for improved patient care.

Keywords: Sickle Cell, Anemia Classification, Teenagers, Bayesian Network, Machine Learning

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

1. INTRODUCTION

Sickle cell anemia, a genetic blood disorder with a high prevalence in teenagers from specific ethnic backgrounds, alters the structure of hemoglobin, impacting the oxygen-carrying capacity of red blood cells [1]. This condition leads to severe health issues, including pain crises, anemia, organ damage, and reduced quality of life. Early diagnosis is paramount [2]. Researchers over the years have shown remarkable contributions in the area of research, to further close the gap, our study proposes enhancing the accuracy of sickle cell anemia classification in teenagers by incorporating additional biomarkers into the dataset.

This augmentation aims to improve upon existing classification systems, with a specific focus on addressing complications in young teenagers. The integration of machine learning and artificial intelligence (AI) in healthcare, particularly Bayesian networks, has channeled in a new phase in disease detection [3]. This study focuses on developing a machine learning system for sickle cell anemia classification in teenagers, utilizing Bayesian networks to model complex relationships within health datasets [4]. By using Bayesian network algorithms, this research aims to construct a predictive model using diverse health data. The model incorporates genetic information, clinical assessments, and patient demographics to discern the likelihood of sickle cell anemia in teenagers, enabling early intervention and personalized treatments [5].

Early detection is crucial for timely medical interventions and improved quality of life. The Bayesian network approach enhances prediction accuracy and provides insights into disease-influencing factors [6]. This study uses Machine Learning to transform sickle cell anemia management, offering more effective and compassionate care for affected teenagers [7]. Sickle cell anemia predominantly affects teenagers of African, Mediterranean, Middle Eastern, and Indian descent [7]. Early detection is vital in regions with a high prevalence of the disease, as it significantly impacts teenagers' educational, social, and emotional well-being [8]. The machine learning system, rooted in the Bayesian network algorithm, considers a comprehensive array of factors influencing the disease's onset and progression, including genetic markers, family history, environmental influences, and clinical symptoms [9]. It is a proactive means of addressing the challenge of sickle cell anemia in teenagers [10].

The study creates a holistic diagnostic tool surpassing traditional single factor approaches [11]. The impact of this research transcends diagnosis, providing hope for affected teenagers, their families, and healthcare providers. By understanding the relationship between genetic predisposition and environmental factors, the research facilitates targeted interventions [12]. The combined efforts of technology and healthcare offers a beacon of hope for a brighter future for teenagers living with sickle cell anemia. By advancing our understanding of this complex genetic

disorder and enhancing diagnostic accuracy, the study marks a significant step forward in the journey toward improved care, enhanced quality of life for teenagers living with sickle cell anemia [13]. Among various age groups affected by SCA, teenagers represent a particularly vulnerable population. During adolescence, significant physiological changes occur, and the burden of SCA can substantially disrupt the lives of affected individuals, impacting their education, social interactions, and overall well-being [14]. Early and accurate diagnosis of SCA in teenagers is crucial for timely interventions and improved disease management [15].

2. METHODS

The method used in this research is Bayesian Network, the method uses the Bayes theorem in its operation. A Bayesian network is a probabilistic graphical model that represents relationships between variables in a system using a directed acyclic graph. Each node in the graph corresponds to a variable, and the edges indicate probabilistic dependencies between the variables. Unlike traditional statistical models, Bayesian networks allow for the explicit representation of uncertainty and the modeling of complex relationships among variables. They provide an intuitive framework for reasoning under uncertainty, making them valuable tools in various fields, including artificial intelligence, healthcare, and finance. Bayesian networks are adept at handling incomplete or uncertain information, enabling practitioners to make informed decisions in situations where uncertainty is inherent in the data. Going further, we will divide this section into two different steps as seen in subsection 2.1, and subsection 2.2:

2.1 Direct Acyclic Graph

In this research, a Direct Acyclic Graph (DAG) serves as a graphical representation showing the relationships among the key variables involved in the classification process. Each variable, such as age, platelet count, mean corpuscular hemoglobin concentration (MCHC), red blood cell count, and packed cell volume, is represented as a node within the DAG. The edges in the DAG are directed and denote the probabilistic dependencies between variables. For instance, an arrow from "age" to "platelet count" indicates that age influences platelet count within the model. Importantly, the DAG maintains an acyclic structure, ensuring there are no loops or circular dependencies. This acyclic nature is fundamental in defining clear and unambiguous probabilistic relationships among the variables. The structure of the DAG reflects conditional independence relationships, signifying that certain variables become conditionally independent once others are observed. This characteristic is particularly relevant in the study of sickle cell anemia classification, where understanding the conditional independence of variables enhances the accuracy of the model.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

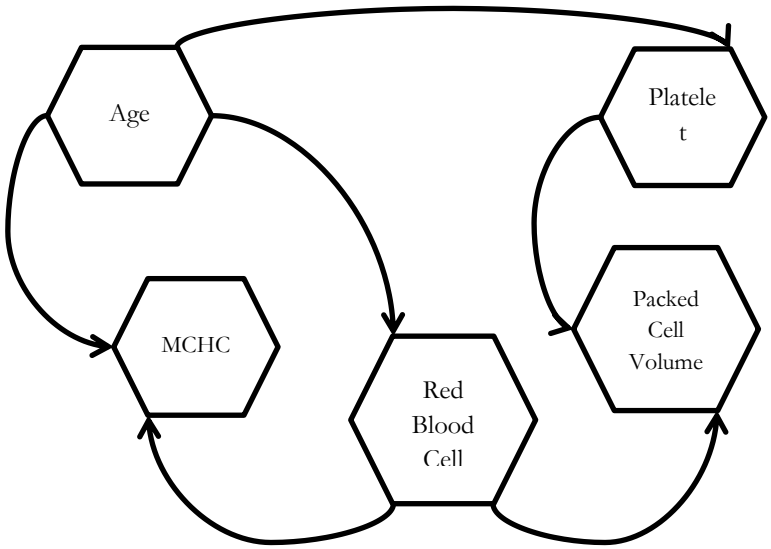

Figure 1. Direct Acyclic Graph of Important Variables

In Figure 1, the Direct Acyclic Graph (DAG) showing the relationships between the various variables is presented. It is a graphical representation of the intricate interplay existing between the variables as they contribute to the classification of sicle cell anemia complication in teenagers. Root nodes, devoid of incoming edges, represent variables that are not influenced by others in the model, while leaf nodes, lacking outgoing edges, signify variables that do not influence other variables. The DAG, through its visual representation, aids in comprehending how information flows through the network during the sickle cell anemia classification process, providing a clear and insightful depiction of the interplay among the medical parameters involved.

2.2 Model Formation

Bayes' Theorem is a mathematical concept in probability theory that calculates the probability of an event A happening given the occurrence of another event B. It is expressed as:

The probability of event A given that event B has occurred (P(A|B)) is equal to the probability of event B given that event A has occurred (P(B|A)) multiplied by the probability of event A occurring (P(A)), divided by the probability of event B occurring (P(B)).

$P(A|B) = (P(B|A) * P(A)) / P(B)$ [8].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

To formulate this problem using Bayesian Networks we will have to establish several conditional and joint probability distributions of the conditional variables in our dataset. This problem is a one Bayesian network problem. We would have one Directed Acyclic Graph for Sickle Cell Anemia model. We will have two classes as we stated earlier in this work, Sickle cell class denoted by (S) and Not Sickle cell class denoted by (N).

To begin with, we wish to model a Bayesian network under the prior hypothesis that a patient is Sickle cell anemia positive (S) given its observed evidence (Z), where Z is the patient data. There is another hypothesis that a patient test is negative (N), still by observing the patient's test tuple or evidence (Z). These hypothesis are presented as likelihoods, such that $P(Z_i|S)$ represents the likelihood of Sickle cell anemia (S) and $(Z_i|N)$ represents the likelihood of the patient not having sickle cell anemia (N), where Z_i for i = n which implies evidence, hence the probability of Sickle cell anemia is P(S) and the probability of its impossibility is P(N) = 1-P(S). Using Bayes Theory, we express our model as follows:

$P(S|Z_i) = P(Z_i|S).\ P(S)/P(Z_i)$. Here $P(S|Z_i)$ is posterior, $P(Z_i|S)$ is likelihood of the patient having sickle cell anemia given Z_i which is the observed evidence on the patient from the test report or medical report and P(N) is the prior probability of No sickle anemia. Where the denominator $P(Z_i)$ being evidence can be expressed as

$$P(Z_i) = P(Z_i|S)P(S) + P(Z_i|N)P(N).$$

For our system, we have two classes namely, class F1 and class F2. Given the dataset row running from $Z = (z_1,z_2,z_3,\ldots,z_n)$ and column running from $A = (k_1,k_2,k_3,\ldots,k_n)$ [3].

Our prior probability may be estimated as $P(S) = D_i/D$. Here D is the total number of training samples and D_i is the total number of training samples of class S.

In Figure 2, the conceptual framework diagram is presented. It shows the processes from the collection of sickle cell anemia dataset, up to the model training, sample classification and Bayesian Network model performance evaluation.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

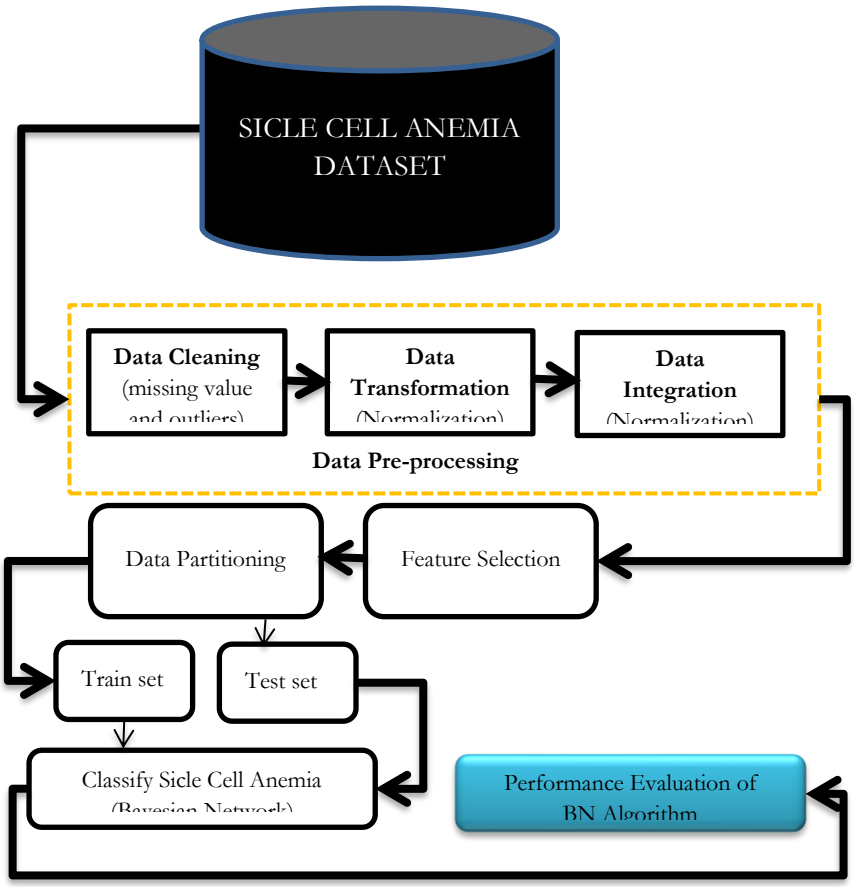

Figure 2. Architecture of the Conceptual Framework

3. RESULTS AND DISCUSSION

In this section, the result of the research is presented. Also, we accompany each result with a corresponding discussion to facilitate understanding and clarity.

Table 1. Classification Report

	Precision	Recall	F1-Score	Support
N(-ve)	0.99	0.99	0.99	160
S(+ve)	0.99	0.99	0.99	150
Accuracy			0.99	315
Macro avg	0.99	0.99	0.99	315
Weighted avg	0.99	0.99	0.99	315

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Model Accuracy: 99.37% In Table 1, the overall performance of the system is 99.40% indicating that the system is at its core in terms of performance on prediction of sickle cell anemia in children.

Confusion Matrix:
[[164 1]
[1 149]]

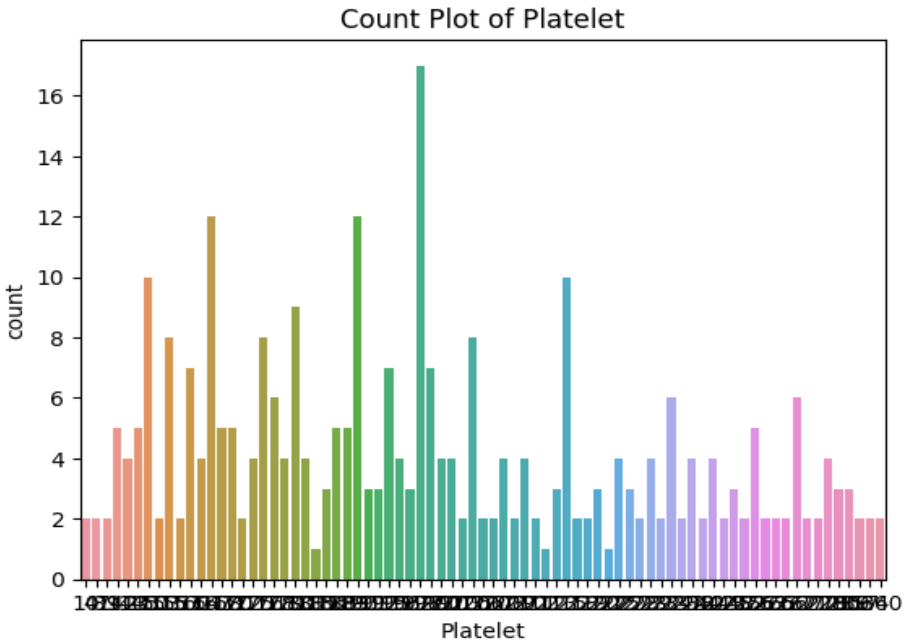

Figure 3. Blood Platelet Count

The bars in Figure 3 shows the different distribution of blood platelet in each of the samples in the dataset used. The unit of measurement for blood platelets is typically expressed as the number of platelets per microliter of blood. The standard unit is often written as "x 10^3/μL" or "thousands per microliter." Platelet count is an essential component of a complete blood count (CBC) and is used to assess the number of platelets in a person's blood. Platelets play a crucial role in blood clotting, and abnormal platelet counts can be indicative of various health conditions.

Journal of Information Systems and Informatics
Vol. 5, No. 4, December 2023
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

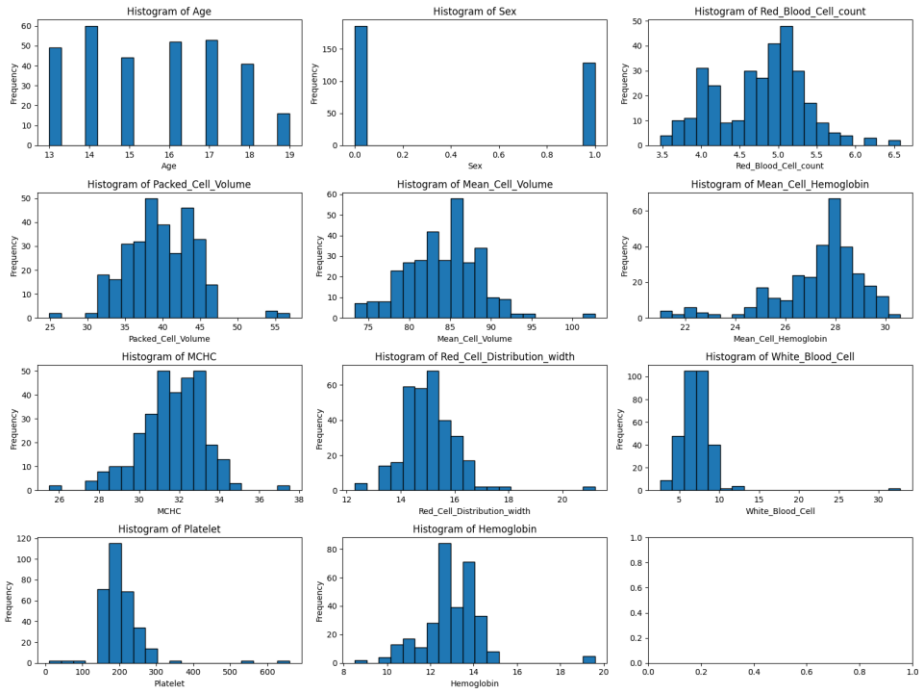

Figure 4. Visualization of Key Features

Figure 4 shows the histogram of the distribution of the different blood components in the sample used. Blood samples, including hemoglobin variants, red blood cell count, and platelet levels, are pivotal in classifying sickle cell anemia. The disease, characterized by abnormal hemoglobin and distorted red blood cells, is identified through the analysis of these parameters. Aberrations in cell count and volume provide insights into the impact of sickle cell anemia on blood composition, while platelet levels offer additional information on vascular health. These quantitative measures enable a comprehensive understanding of the physiological changes induced by sickle cell anemia, aiding accurate and early classification for timely intervention.

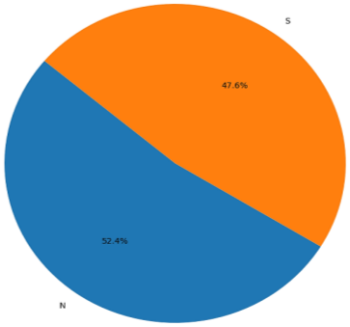

Figure 5. Pie Chart of Sickle Cell Anemia Classes

Journal of Information Systems and Informatics
Vol. 5, No. 4, December 2023
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 5 shows how the two classified classes are distributed in the entire dataset. In this research, the Sickle Cell Anemia positive class represents 47.6% of the cases, while the No Sickle Cell Anemia class comprises 52.4%. The relatively close distribution of these classes, with a slight majority in the No Sickle Cell Anemia class, suggests a balanced representation in the dataset. The societal implication of this balance is significant, as it indicates a realistic reflection of the prevalence of sickle cell anemia within the studied population.

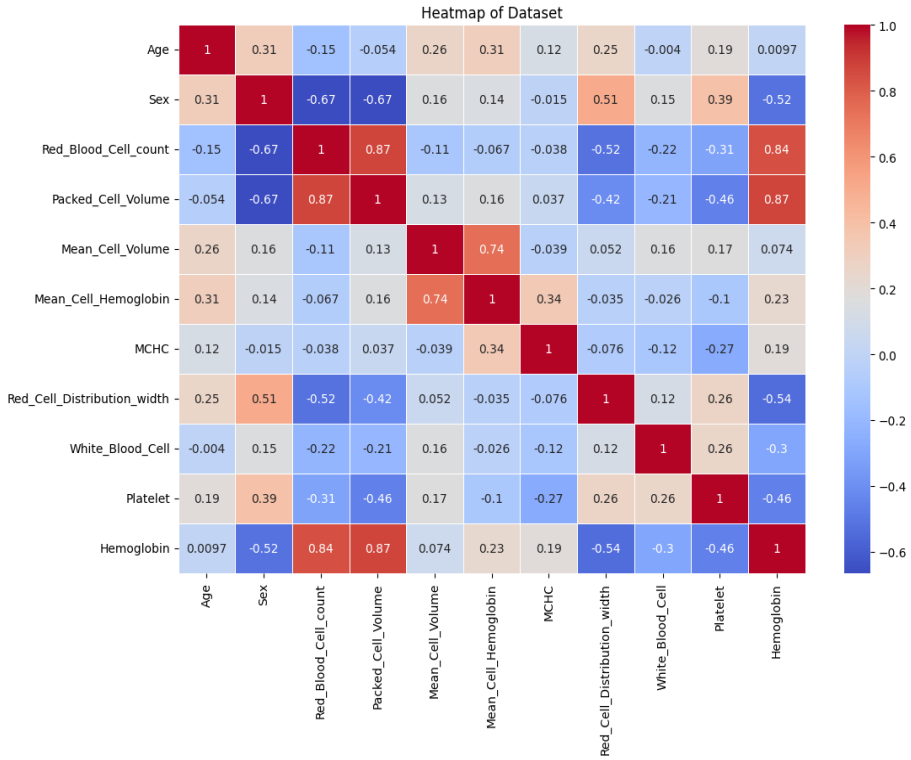

Figure 6. Heatmap Plot of all the Feature

Figure 6 indicates correlation between the variables in the dataset as they contribute to the classification of Sickle Cell Anemia in teens. It is employed to illustrate the strength or weight of connections between variables in the Bayesian network model. Specifically, the heatmap plot represents the conditional probabilities and influence of one variable on another within the Bayesian network. For instance, darker or more intense colors indicate stronger influences, while lighter colors suggest weaker or negligible connections.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

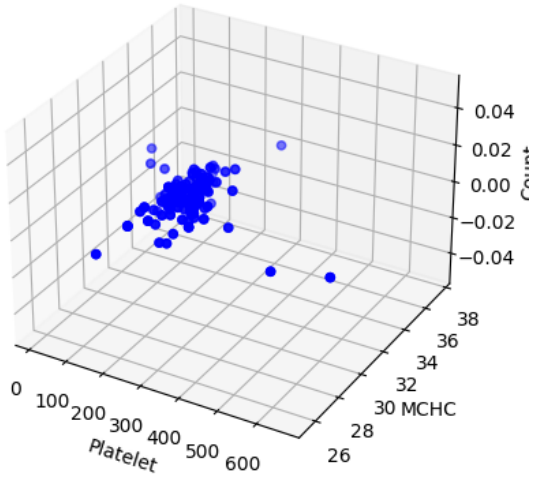

Figure 7. Scatter Plot of Platelet and MCHC

Figure 7 shows the distribution of blood plateet and MCHC in the dataset. Platelets, or thrombocytes, are small cell fragments that play a crucial role in blood clotting (hemostasis). They are involved in the formation of blood clots to prevent excessive bleeding. In sickle cell anemia, individuals may experience complications related to blood vessel occlusion, which can sometimes lead to a higher risk of clot formation. MCHC (Mean Corpuscular Hemoglobin Concentration), is a measure of the concentration of hemoglobin in a given volume of packed red blood cells. It is usually expressed as a percentage. In sickle cell anemia, the main problem lies in the abnormal hemoglobin (hemoglobin S) that causes red blood cells to take on a characteristic sickle shape.

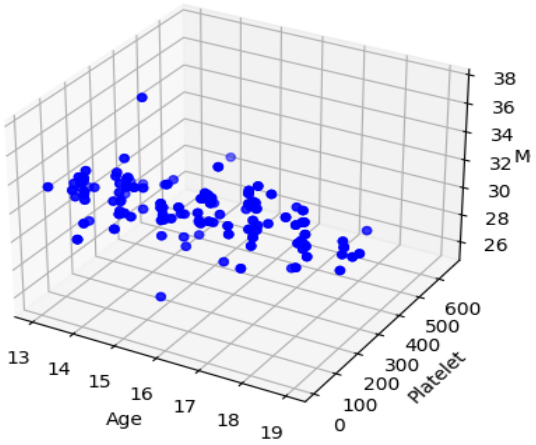

Figure 8. Distribution of Age and Platelet

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 8 shows how platelet changes across ages. Blood platelet increases as one leaves teenage age. Sickle cell anemia is a genetic disorder present from birth. Diagnosis usually occurs through newborn screening or later in life if symptoms or complications arise. Platelet count may be monitored as part of the overall health assessment. In sickle cell disease, there is an increased risk of certain complications, including clotting events. Platelet count can provide information about the risk of thrombosis or other vascular complications. MCHC is a measure of hemoglobin concentration in red blood cells. Changes in MCHC can reflect overall changes in the composition of blood cells. Sickle cell anemia is characterized by the presence of abnormal hemoglobin (hemoglobin S), leading to the distinctive sickle shape of red blood cells.

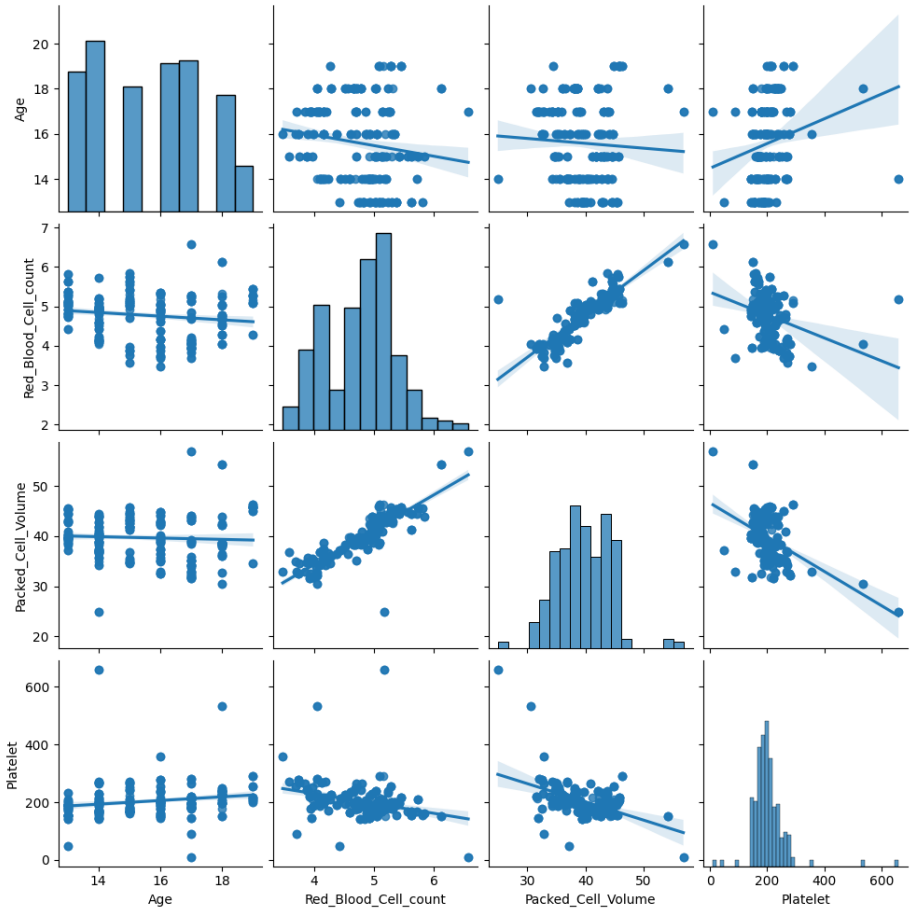

Figure 9. Regression Plot of Age, Red_Blood_Cell_count, Packed_Cell_Volume and Platelet

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 9 shows the distribution of the four features in the dataset and how they contribute to the formation and detection of complication of Sickle Cell Anemia. Red blood cell count (RBC) and packed cell volume (PCV) are two additional hematological parameters that are relevant in the context of sickle cell anemia. Red blood cell count is a measurement of the number of red blood cells in a given volume of blood. In sickle cell anemia, the RBC count might be elevated (a condition known as polycythemia) because the body attempts to compensate for the reduced oxygen-carrying capacity of the sickle-shaped red blood cells. PCV measures the proportion of blood that is occupied by red blood cells. It is expressed as a percentage. Similar to RBC count, an elevated PCV may be observed in teenagers with sickle cell anemia. This is again a compensatory mechanism to maintain oxygen delivery, given the altered structure and function of sickle cells. Both RBC count and PCV are part of the complete blood count (CBC), a routine blood test that provides information about various blood components. These paramters are key in the classification of sickle cell anemia complication in teenagers.

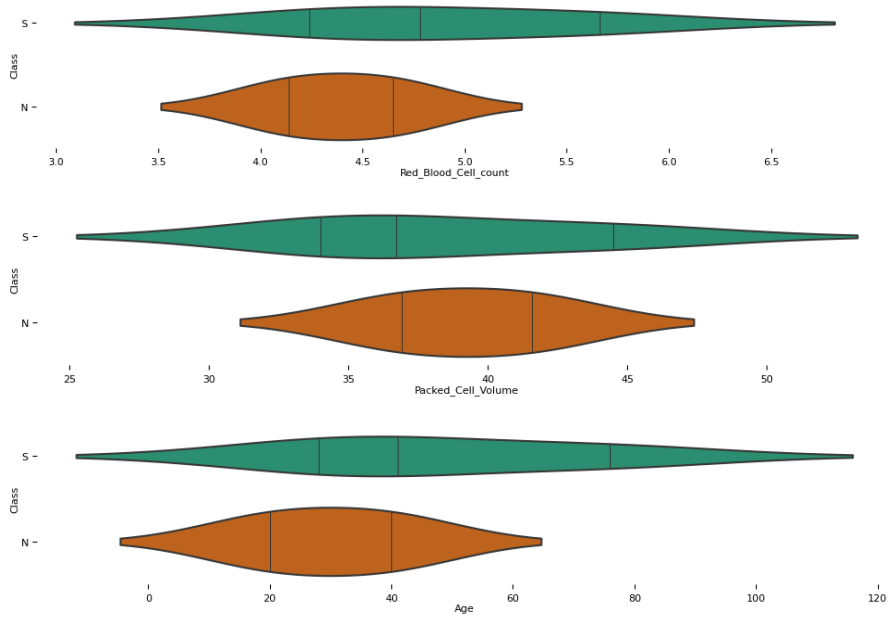

Figure 10. Showing Categorical Features

Figure 10 shows that the relationship between packed cell volume (PCV), red blood cell count (RBC), and age is complex and can be influenced by various factors, including the presence of medical conditions like sickle cell anemia. In general, PCV levels can vary with age. For example, newborns and infants typically

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

have higher PCV values, and these values gradually decrease as a child grows. Adult levels are generally reached by adolescence. However, in sickle cell anemia, the relationship can be altered. Sickle cell anemia often leads to chronic hemolysis (destruction of red blood cells), and as a compensatory mechanism, the body may increase the production of red blood cells, which could result in an elevated PCV. Similar to PCV, the relationship between RBC count and age can be influenced by factors such as growth and development. In sickle cell anemia, the RBC count might be elevated due to increased production as a response to the hemolysis of sickle-shaped red blood cells. Sickle cell anemia is a genetic condition, and its symptoms can manifest early in life. The relationship between age and the impact of sickle cell anemia on blood parameters is more directly related to the progression of the disease rather than a typical aging process. In teenagers with sickle cell anemia, you might observe changes in blood parameters that reflect the ongoing challenges associated with the disease, such as an increased RBC count and PCV.

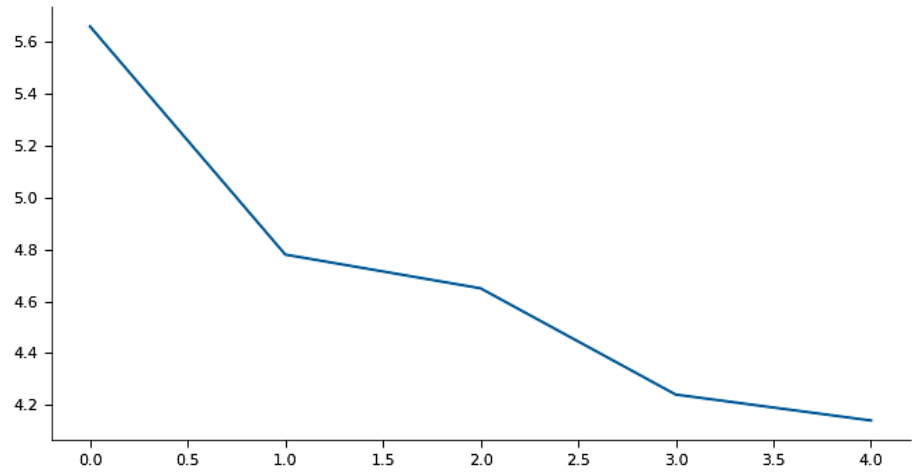

Figure 11. Red Blood Cell Count Distribution

Figure 11 indicates that the red blood cell count (RBC count) is an important component of a complete blood count (CBC) and provides valuable information about the number of red blood cells in a person's bloodstream. In the context of sickle cell anemia in teenage ages, the RBC count becomes particularly significant. Red blood cells (RBCs) contain hemoglobin, a protein that binds with oxygen and carries it from the lungs to the rest of the body. In sickle cell anemia, the structure of hemoglobin is abnormal, leading to the characteristic sickle shape of red blood cells. These abnormal cells can clump together, leading to reduced oxygen-carrying capacity. Monitoring the RBC count helps assess the overall oxygen-carrying capacity of the blood. Sickle cell anemia is a type of hemolytic anemia, meaning that red blood cells are destroyed more rapidly than the body can replace them.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

This can result in a lower-than-normal RBC count. Anemia can contribute to fatigue, weakness, and other symptoms. The RBC count is a key parameter in diagnosing and monitoring the severity of anemia in individuals with sickle cell disease. Figure 12 indicates that in the whole sample collected, class S had the highest frequency of appearance compared to class N.

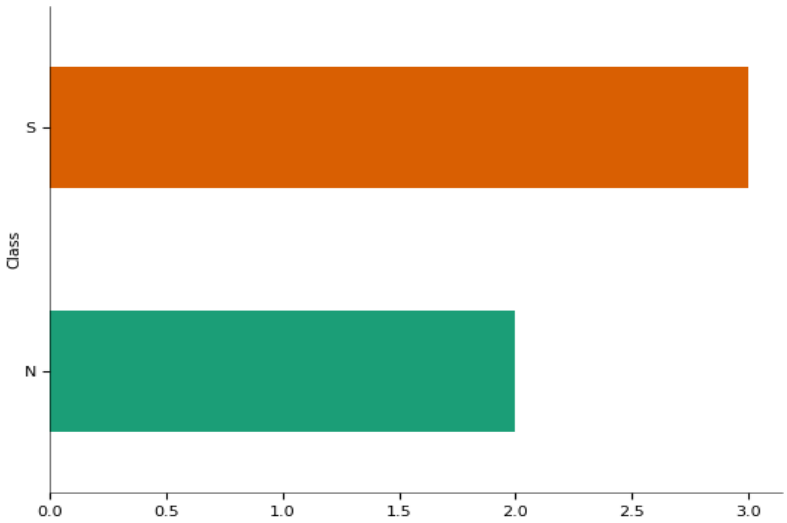

Figure 12. Bar Chart of Class S and N

4. CONCLUSION

In this study, our primary focus was on classifying the complications of sickle cell anemia specifically in teenagers. The dataset utilized comprises medical diagnostic information pertaining to young individuals aged between 13 and 19 years. The emphasis on this age group is intentional, as we sought to gain insights into the unique challenges faced by teens dealing with sickle cell issues. Our findings revealed two distinct groups among teenagers grappling with sickle cell complications. The first group consists of individuals who are not entirely positive for sickle cell anemia but are impacted to some extent due to a gene mutation. These teens exhibit the AS genotype, where the normal gene factor in the bloodstream is denoted by 'A,' while the sickle cell factor is represented by 'S.' Remarkably, this group may experience relatively better health compared to the second group, given the presence of the normal gene factor.

Conversely, the second group comprises individuals with the SS genotype, commonly known as "sicklers." This group faces frequent bouts of illness and is at a heightened risk of fatality. Their vulnerability stems from the absence of a normal gene factor, leaving them susceptible to severe complications even with

minor illnesses. Through our classification, we have categorized sickle cell anemia in teenagers into two classes: Class S, indicating a positive status, and Class N, representing a negative status. It's noteworthy that our dataset exclusively includes teenagers, a deliberate choice aimed at ensuring the effectiveness and accuracy of our classification system within this specific age group. Our goal is to contribute to a more understanding of sickle cell anemia and its implications for teenagers, ultimately paving the way for targeted interventions and improved healthcare outcomes. In conclusion, we wish to first state that sickle cell anemia condition does not change as a child grows older, as it is speculated in the society. The condition is a lifelong situation. However, those with AS have been discovered to live longer than those with SS.

REFERENCES

[1] D. N. Mukund, G. S. Shailesh, and A. Rajanikanth, "A Brief Bibliometric Survey of Leukemia Detection by Machine Learning and Deep Learning Approaches," *Library Philosophy and Practice*, 2020.

[2] Z. Huang, J. Lin, L. Xu, H. Wang, T. Bai, Y. Pang, and T. H. Meen, "Fusion High-Resolution Network for Diagnosing Chest X-ray Images," *Electronics*, vol. 9, 190, 2020.

[3] Ekong, B. Ekong, and A. E. Edet, "Supervised Machine Learning Model for Effective Classification of Patients with Covid-19 Symptoms Based on Bayesian Belief Network," *Researchers Journal of Science and Technology*, vol. 2, pp. 27-33, 2022.

[4] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, and Y. Duan, "Deep Learning Models Classification of Red Blood Cells in Microscopy Images to Aid in Sickle Cell Anemia Diagnosis," *Electronics*, vol. 9, no. 427, 2020. doi:10.3390/electronics9030427.

[5] A. Elsabagh et al., "Artificial Intelligence in Sickle Disease," *Elsevier*, pp. 2-9, 2023.

[6] U. J. Devi, M. Devaki, S. Bhusal, and V. A. Konda, "Anemia Detection using Machine Learning," *International Journal of Research Publication and Reviews*, vol. 4, no. 4, pp. 1996-2011, 2023.

[7] K. J. Wenger, C. E. Koldijk, E. Hattingen, L. Porto, and W. Kurre, "Characterization of MRI White Matter Signal Abnormalities in the Pediatric Population," *Children*, 2023.

[8] E. Edet and G. O. Ansa, "Machine Learning Enabled System for Efficient Classification of Intrusion Severity," *Global Journal of Engineering and Technology Advances*, vol. 16, no. 3, pp. 41-50, 2023.

[9] Albayrak et al., "Sickle Cell Anemia Detection," *Medical Technologies National Congress (TIPTEKNO)*, pp. 1-4, 2018.

[10] T. S. Chy and M. A. Rahaman, "Automatic Sickle Cell Anemia Detection Using Image Processing Technique," *in Proc. 2018 IEEE International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pp. 1-4, 2018.
[11] L. Alzubaidi et al., "Classification of Red Blood Cells in Sickle Cell Anemia Using Deep Convolutional Neural Network," *in International Conference on Intelligent Systems Design and Applications*, pp. 550-559, 2018.
[12] B. Fang, Y. Lu, Z. Zhou, Z. Li, Y. Yan, L. Yang, G. Jiao, and G. Li, "Classification of Genetically Identical Left and Right Irises Using a Convolutional Neural Network," *Electronics*, vol. 8, 1109, 2019.
[13] V. Acharya and K. Prakasha, "Computer-Aided Technique to Separate the Red Blood Cells, categorize them and Diagnose Sickle Cell Anemia," *Journal of Engineering Science Technology Review*, vol. 12, no. 2, 2019.
[14] Z. Huang, J. Lin, L. Xu, H. Wang, T. Bai, Y. Pang, and T. H. Meen, "Fusion High-Resolution Network for Diagnosing ChestX-ray Images," *Electronics*, vol. 9, 190, 2020.
[15] S. Nurmaini, A. Darmawahyuni, M. Sakti, M. N. Rachmatullah, F. Firdaus, and B. Tutuko, "Deep Learning-Based Stacked Denoising and Autoencoder for ECG Heartbeat Classification," *Electronics*, vol. 9, 135, 2020.