Utilizing Machine Learning to Identify High-Risk Groups in Sickle Cell Anemia

A Single-Center Experience in Saudi Arabia

Haneen Banjar[1], Nofe Alganmi[2], Hajar Alharbi[3], Ahmed Barefah[4], Hatem Alahwal[5], Salwa Alnajjar[6], Abdulrahman Alboog[7], Salem Bahashwan[8], Galila Zaher[9]

Department of Computer Science-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2, 3]
The Center of Research Excellence in Artificial Intelligence and Data Science,
King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2]
Centre of Artificial Intelligence in Precision Medicines, King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2]
Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2]
Faculty of Medicine, King Abdulaziz University, Jeddah 21589, Saudi Arabia[4, 5, 6, 8, 9]
Faculty of Medicine, University of Jeddah, Jeddah, Saudi Arabia[7]

***Abstract*—Sickle Cell Anemia (SCA) is a hereditary condition causing abnormal red blood cells, leading to severe health complications. Traditional treatment approaches for SCD often involve reactive management, which can delay appropriate interventions and worsen patient outcomes. The aim of this study is to leverage machine learning (ML) algorithms, including Logistic Regression (LR), Support Vector Machines (SVM), and Decision Trees (DT), to identify high-risk groups among SCA patients using clinical and pathological data from King Abdulaziz University Hospital. This study employs a comprehensive dataset comprising 200 SCA patients, with data preprocessing to handle missing values and feature selection techniques to enhance model performance. The dataset is divided into training and testing sets, and models are evaluated using ten-fold cross-validation. Performance metrics such as True Positive Rate (TPR), False Negative Rate (FNR), Positive Predictive Value (PPV), and False Discovery Rate (FDR) are used to assess model effectiveness. The results indicate that the SVM model with the top seven correlated features achieved the highest TPR and PPV, along with the lowest FNR and FDR, demonstrating its superior performance in identifying high-risk patients. The study concludes that ML models, particularly SVM, can significantly improve risk assessment and patient management in SCA, offering a proactive tool for healthcare providers. The main message is the potential of ML algorithms to enhance clinical decision-making and improve outcomes for patients with SCA.**

Keywords*—Sickle cells anemia; feature selection; predicting complication; machine learning

I. Introduction

SCD is a hereditary condition that results in the production of abnormal red blood cells. The disease was first identified in 1910 [2]. Today, it is estimated that approximately 5% of the global population carries a gene for sickle cell disease or thalassemia [1]. Annually, over 300,000 newborns are affected by severe forms of these disorders, with most cases occurring in low- and middle-income countries [3], [4]. SCD is particularly prevalent in regions such as the Mediterranean, Africa, India, the Caribbean, and the Middle East.

In Saudi Arabia [5], SCD is a common genetic disorder, with carrier rates varying from 1.4% to 2% in certain areas and reaching as high as 27% in some regions. The Saudi premarital screening program has revealed that 0.26% of the adult population is affected by SCD, while 4.2% carry the sickle cell trait. The highest prevalence is observed in the Eastern province, with SCD affecting about 1.2% of the population and sickle cell trait carriers comprising approximately 17%. Current treatment decision-making for SCD involves monitoring symptoms and complications and adjusting treatment plans accordingly. This reactive approach can significantly impact patients' lives, leading to delayed treatment options and irreversible disease complications. Despite extensive epidemiological surveillance, no publicly reported model has yet been validated on a Saudi cohort to predict severe events; this gap leaves clinicians reliant on qualitative heuristics and underscores the need for data-driven, localised risk stratification.

Sickle-cell complications place a persistent burden on emergency and hematology services in Saudi Arabia; delayed therapeutic escalation is common because current risk assessments rely on clinician intuition or post-event markers. By embedding a data-driven prognostic layer into routine complete-blood-count (CBC) workflows, clinicians can transition from reactive to proactive care. Applying artificial intelligence (AI) tools, particularly Machine Learning (ML) algorithms, to predict the severity and complications of SCD in diagnosed patients, offers a proactive solution that could assist physicians in making more informed treatment decisions. ML techniques can analyze vast amounts of clinical and pathological data to identify patterns and risk factors that are not easily discernible through traditional methods. Consequently, the objective of this study is to leverage ML algorithms, including Logistic Regression, Support Vector Machines, and Decision Trees, to identify high-risk groups among SCD patients using data from King Abdulaziz University Hospital. By developing predictive models, we aim to enhance the accuracy of risk assessments and improve patient management. Directly addresses a national patient-safety gap and aligns with Vision 2030's precision-

medicine goals. Research Question (RQ): Which of the evaluated machine-learning algorithms (Logistic Regression, Support Vector Machine, or Decision Tree) most accurately identifies high-risk SCA patients using routinely collected hematological features in a Saudi clinical setting? The remainder of this paper is organised as follows: Section II details the dataset and pre-processing pipeline. It describes the three candidate algorithms and feature-selection procedure; Section III sets out the experimental protocol. It presents quantitative results; Section IV discusses clinical implications and algorithm–data suitability; finally, Section V concludes with limitations and future work.

A. Related Work

SCD is a complex genetic disorder characterized by significant phenotypic variability [6]. The development of accurate prediction models for SCD severity and outcomes is crucial for improving patient care and treatment decision-making [7]. ML techniques have shown promise in interpreting medical data and predicting disease severity, complications, and treatment dosages [7]. Bayesian network modeling has been used to create a personalized disease severity score that predicts the risk of death within 5 years, incorporating factors such as renal insufficiency, leukocytosis, and hemolytic anemia severity [8]. While known modifiers like fetal hemoglobin levels and α-thalassemia influence disease severity, other genetic variants and environmental factors, including climate and air quality, may also play a role [6]. Despite recent successes, modeling the multisystem pathology of SCD remains challenging, and further research is needed to improve prediction of specific adverse outcomes and global disease severity [9].

Research on predicting disease severity in SCD has focused on various factors and approaches. A systematic review identified multiple indices used to assess SCD severity, incorporating elements like organ damage and complications, but found a lack of consensus and validation for these measures [10]. To address this, experts developed a severity classification system using a modified Delphi panel, considering factors such as age, genotype, and organ damage [11]. ML techniques have shown promise in predicting acute organ failure in critically ill SCD patients, using physiological markers derived from vital signs [12]. Despite these advancements, modelling the variable and multisystem pathology of SCD remains challenging. While there have been some successes in predicting specific adverse outcomes and global disease severity, significant challenges persist in developing comprehensive prognostic tools for SCD [9].

Feature selection techniques are crucial for improving model performance in clinical prediction by identifying relevant features from large datasets [13]. Common approaches include filter and wrapper methods [14]. Filter methods use criteria like correlation, significance, or variable importance to rank features, while wrapper methods employ algorithms to evaluate feature subsets [13]. Hybrid approaches combining filter and wrapper techniques have shown promise in generating minimal redundancy maximal relevant feature subsets [15]. A tri-stage wrapper-filter framework has been proposed, utilizing ensemble filter methods, correlation analysis, and meta-heuristic optimization to select optimal features for disease classification [16]. These techniques aim to reduce dimensionality, remove irrelevant or redundant data, and improve classification accuracy [16], [14]. Effective feature selection can lead to more parsimonious models, reduced training time, and improved predictive performance in clinical settings [13], [16].

ML algorithms have shown significant potential in improving healthcare outcomes for patients with SCD. Studies have demonstrated that ML models outperform traditional readmission risk scoring systems in predicting 30-day hospital readmissions for SCD patients [17]. Random Forest algorithms have been particularly effective, achieving higher accuracy and area under the curve (AUC) scores compared to standard [17]. ML techniques have also been successful in predicting SCD from erythrocyte smears, with ensemble models like Random Forest-XGBoost showing superior performance [18]. The integration of ML in healthcare extends beyond SCD, offering opportunities for personalized treatment and improved resource allocation across various medical conditions [19]. These advancements in ML-driven predictive models have the potential to revolutionize healthcare systems, leading to more efficient and patient-centered care.

II. METHODS

A. Patient Population

The 200 patients from King Abdulaziz University Hospital (KAUH) were eligible for the study, and any patients who were not diagnosed with SCA were excluded. The dataset retrieved data from electronic health records, including each patient's complications records and clinical and pathological data. The data were accessed for research purposes from (1/10/2020). All patients' identities were anonymized, and authors did not have access to information that could identify individual participants during or after data collection. Table I shows the features that were included in the study. Ethical approval was obtained from the unit of the Biomedical Ethics Research Committee at KAUH (Reference No. 511-20). As participants' data were retrospectively included from medical records, the need for informed consent was waived by the ethics committee, as the data were fully anonymized before access.

B. Patients Outcomes

The patients enrolled in the study were divided into two broad outcome groups: (i) Low-risk group: SCA patients who did not develop any complications; and (ii) high-risk group: SCA patients who developed complications. The complications include: Vaso-Occlusive Crisis (VOC), Hepatorenal, gallstone, Stroke, Osteotomy, Avascular Necrosis (AVN), Chronic Kidney Disease (CKD), acute tubular necrosis, Upper Respiratory Tract Infection (URTI), cute Chest Syndrome (ACS), pulmonary Embolism (PE), Pneumonia, Cholecystitis, and Priapism).

C. Datasets

The dataset used in this study comprises three distinct sheets: "ContWithMissing", "ContNoMissing," and "CatWithNoMissing". Each sheet presents a different aspect of the data, catering to specific requirements for ML model development and evaluation. Firstly, the "ContWithMissing" sheet contains continuous data with some values missing, indicating a need for imputation or exclusion during preprocessing. This sheet is crucial for understanding data

completeness and planning subsequent data cleaning steps. Secondly, the "ContNoMissing" sheet provides continuous data with no missing values. This dataset guarantees that all continuous data inputs are available for analysis. Thirdly, the "CatNoMissing" sheet contains categorical data with all missing values addressed. The number of rows reduced compared to the original dataset because some Thalassemia patients were excluded. This sheet ensures that categorical data inputs are fully available for analysis and modeling.

TABLE I. CLINICAL AND PATHOLOGICAL FEATURES

Features	Description
Age	Clinical feature recorded at the time of diagnosis.
Gender	Clinical feature recorded at the time of diagnosis.
Hemoglobin level (Hb)	An internal protein within red blood cells that serves as a transport system for oxygen from the lungs to the tissues and organs of the body, and a transport system for carbon dioxide from the body to the lungs.
White Blood Cell (WBC)	A cell in the blood and lymph is produced in the bone marrow. White blood cells are a component of the body's immune system. They assist the body in fighting off illness and disease. Granulocytes (neutrophils, eosinophils, and basophils), monocytes, and lymphocytes are the different kinds of WBCs (T cells and B cells).
Platelets	Large cells in the bone marrow known as megakaryocytes form platelets. They assist in forming blood clots that can slow or halt bleeding and assist in wound healing.
Mean Corpuscular Volume (MCV)	The MCV blood test evaluates the erythrocyte count, which is the average size of red blood cells.
Mean Cell Hemoglobin (MCH)	The typical hemoglobin level in a typical red blood cell. The MCH is calculated by calculating the red blood cell count and hemoglobin levels.
Mean Cell Hemoglobin Concentration (MCHC)	The hemoglobin concentration in red blood cells, on average.
Reticulocyte	Newly produced, immature red blood cells, known as reticulocytes, emerge from the bone marrow. They form and ripen in the bone marrow before they are released into the blood.
Red Cell Distribution Width (RDW)	A red cell distribution width (RDW) test is used to determine your blood's red blood cell volume and size range (erythrocytes).
Hemoglobin Electrophoresis: Hb A Hb A2 Hb F Hb S	Electrophoresis for hemoglobin is a procedure that uses a high-voltage electric field to separate the various types of hemoglobin found in the blood. Additionally, it searches for abnormal forms of hemoglobin. The most commonly occurring type of hemoglobin (Hg A) in healthy adults. Hb F Fetal hemoglobin. Unborn babies and newborns have a specific type of hemoglobin. Immediately after birth, Hb F is replaced by Hb A. Sickle cell disease is the common cause of hemoglobin of this type.
Aspartate Amino Transferase (AST)	A special enzyme, present in increased levels in the blood following a heart attack or liver disease, aids in the transfer of an amino group from glutamic acid to oxaloacetic acid.
Alanine Amino Transferase (ALTI)	A naturally occurring enzyme found in the liver and heart. The ALT enzyme is released into the blood when liver or heart injuries occur. As a result of liver damage (such as from viral hepatitis) or an insult to the heart, the blood ALT levels increase.
Alkaline phosphatase (ALPI)	An enzyme which is commonly found in various body parts, but is most prevalent in the liver, bones, intestine, and kidneys. A blood test to measure the activity of alkaline phosphatase is used to gauge the enzyme's concentration. Elevated ALP levels may be a result of liver disease or another type of health problem.
Gamma-glutamyl Transferase (GGT)	An enzyme found throughout the body in many organs at the highest concentrations found in the liver. Elevated GGT can be found in the blood of most people with liver or bile duct problems. The blood sample test is meant to gauge the level of GGT. When the liver is injured, GGT levels increase normally, but if the liver is compromised, they will be higher.

D. Framework

The framework (Fig. 1) is divided into three sections: data preprocessing, features selection techniques, and ML methods.

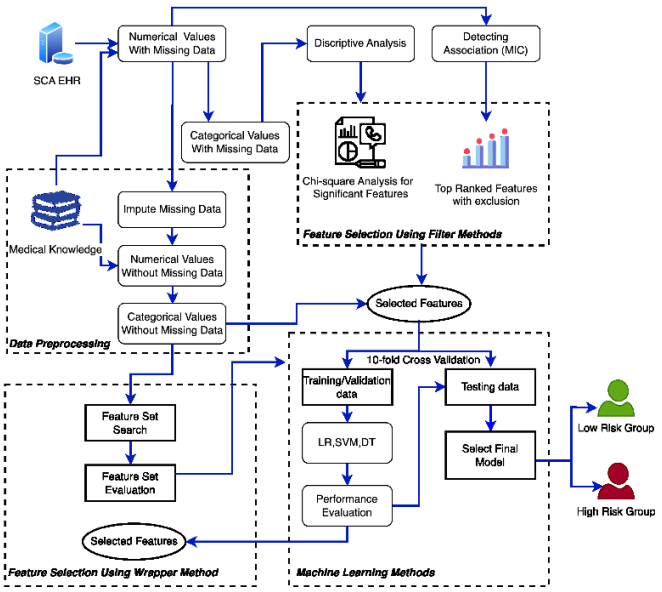

Fig. 1. Framework for preprocessing, feature-selection and modelling.

1) Data preprocessing: The dataset included numerical values with missing data. This step included: imputing missing data and reformatting the data using medical knowledge.

2) Impute missing data: Two of the most frequently used approaches for dealing with missing data are imputation and deletion, and the former can introduce bias while the latter introduces both bias and a reduction in statistical power [20]. In this step, linear interpolation uses the two nearest points to estimate the value of the interpolated point in a one-dimensional data sequence [21]. For all features, the numerical values were imputed with the linear interpolation using SPSS 21 since removing patient data reduces patient datasets, which can affect the performance of learned models, we chose to eliminate those features with missing percentages above 10%.

3) Reformate the features using medical knowledge: We used existing knowledge, including standard blood count boundaries [22], and clinical expertise domains, to format data into categories.

4) Feature selection techniques: Feature selection algorithms are broadly divided into three categories: filter

methods, wrapper methods, and embedded method [23]. Here we first used filter methods.

A statistical score is applied to each feature through the use of filter feature selection methods. The results are rated by score, and then the preferred features are selected to be retained or discarded from the dataset. The methods are typically univariate and either treat the feature in isolation or regard it in the context of the dependent variable. Here, the top correlated features include finding new patterns in large datasets, which highlight interesting relationships between pairs of variables. For two-variable relationships, we used two techniques as a measure of dependence: i) the maximal information coefficient (MIC)[24]. The results were sorted from largest to lowest and discarded as the lowest score feature, ii) the Chi-squared test which is a nonparametric statistical test that is widely used to assess the statistical significance of predictive factors [25]. We used Recursive Feature Elimination (RFE) which iteratively builds models and removes the least key features.

The second feature selection technique was the wrapper method which evaluates the performance of a subset of features by training a machine learning model. They use search algorithms to find the best subset of features based on the model's performance. Two common methods are known as forward selection and backward elimination. We used forward selection to start with an empty set and added features one by one based on performance improvement.

This study used hybrid feature selection approach that combines the strengths of both methods (filter and wrapper). Initially, features are ranked using specific criteria, followed by applying a wrapper algorithm to generate a subset from the ranked features [15].

We chose to use filter and wrapper methods for feature selection instead of embedded methods because filter methods are computationally efficient and model-agnostic, allowing for quick preliminary feature evaluation. Wrapper methods, though more computationally intensive, account for feature interactions and improve model performance by evaluating subsets through model training. In contrast, embedded methods are model-dependent, which could limit the flexibility and generalizability of our feature selection process. By using filter and wrapper methods, we ensured a versatile and robust feature selection applicable across various machine learning models.

5) Machine learning methods: Three ML methods were utilized: Logistic regression (LR), Support Vector Machine (SVM), and Decision Tree (DT). LR is a statistical method used to model the probability of a binary outcome based on one or more predictor variables [26]. It is particularly useful in scenarios where the dependent variable is categorical, such as pass/fail or disease/no disease [27]. Unlike linear regression, which predicts continuous outcomes, logistic regression estimates the probability of an event occurring versus not occurring [27]. LR implemented using the fitclinear function in MATLAB 2024a. In the context of this study, LR is employed to predict whether patients fall into high-risk or low-risk categories based on their features. The fitclinear function optimizes the model parameters using maximum likelihood estimation to find the best fit for the data. The logistic function (sigmoid) transforms the linear output into a probability, which is then used to make binary predictions. The model's performance is evaluated using cross-validation to ensure robustness and avoid overfitting.

Another common ML model is SVM which is a supervised learning method used for both regression and also classification [28].However, they are most commonly utilized for binary classification. The SVM algorithm aims to find an appropriate hyperplane in such an N-dimensional space that accurately classifies the data points [29]. The memory efficiency of SVM and its efficiency for high-dimensional environments are the two key benefits of employing it. SVMs are among the most extensively used and successful ML algorithms in supervised learning for classifying and recognizing problems. They get a solid theoretical foundation, which makes them invaluable throughout this field. The essential principle behind SVMs was that with some high input vector x and vector output y, there must be some unknown and non-linear relationship (mapping, function) $y = f(x)$. There seems to be no knowledge about input data vectors' underlying joint distributions. The training data would be the only information provided. As a result, they are classified as supervised learning algorithms. SVMs create a hyperplane to divide two classes. The algorithm intends to attain the greatest possible separation between the classes.

SVM finds the hyper-plane as in feature space which distinguishes between categories for classification. A training data point has been represented as points within feature space by an SVM model, which is mapped in a way in which points belonging to the different classes were also separated by just as wide a margin as possible. The testing data points then are mapped into the same space and classified according to where they fall upon that margin. SVM implemented here using the fitcsvm function in MATLAB 2024a. In this study, SVM is utilized to classify patients into high-risk and low-risk categories based on their features.

The third ML method is DT, which is a popular machine learning method used for classification tasks. DTs are effective in handling non-linear relationships and can be scaled for big data applications [30]. DTs have been applied in various fields, including education, building, botany, social science, and medicine [31]. DT implemented using the fitctree function in MATLAB 2024a, the DT functions by repeatedly dividing the dataset into smaller subsets according to the values of input features, creating a structure that resembles a tree. In this tree, each node corresponds to a feature, each branch signifies a decision rule, and each leaf node indicates an outcome. The fitctree function grows the tree by selecting features that best split the data according to a chosen criterion, such as Gini impurity or entropy.

The process of building a decision tree begins with selecting the best feature for the root node. The algorithm evaluates all features and selects the one that best separates the data into distinct classes. This is done by calculating the split criterion for each feature. Gini impurity was used to measure the probability of a randomly chosen element being misclassified if it was randomly labeled according to the distribution of labels in the subset. Lower values of Gini impurity indicate better splits.

Once the best feature is chosen, the data is split into subsets based on the values of this feature. For categorical features, this involves creating branches for each category. The process of selecting the best feature and splitting the data is repeated recursively for each subset. At each node, the algorithm considers only the subset of data that reaches that node. The recursion continues until one of the stopping criteria is met: all instances in a node belong to the same class, there are no remaining features to split on, the maximum tree depth is reached, or a minimum number of instances per node is specified.

In this study, the DT model is used to classify patients into high-risk and low-risk categories based on their features. The model is trained using 90% of the dataset and validated using 10% of the dataset. Ten-fold cross-validation is applied to ensure the model's robustness and to prevent overfitting.

The selection of the most appropriate machine learning method depends on the specific application and dataset characteristics [31].

6) Performance evaluation: By using the confusion matrix, one can see how the selected model performed in each category. The confusion matrix helps you figure out where the model messed up. The number of observations is defined as the following:

- TP: patients are correctly categorized as low risk.
- FP: patients categorized in low risk, but they predicted as high risk.
- TN: patients are correctly categorized as high risk.
- FN: patients categorized in high risk, but they predicted as low risk.

Final Model and Feature Set Selection: For the final model and feature set selection, we employed a comprehensive evaluation process based on four key performance metrics: True Positive Rate (TPR), False Negative Rate (FNR), Positive Predictive Value (PPV), and False Discovery Rate (FDR). The TPR is the ratio of true high-risk classifications to true high-risk cases. The FNR is the proportion of incorrectly classified instances in relation to the number of actual classes in high risk. PPV stands for predicted in a high risk to correctly classified in high-risk observation ratio. The FDR is the rate of predictions that are incorrect per class that was classified as high risk.

Initially, we applied filter methods to efficiently evaluate and reduce the dimensionality of the dataset, using statistical measures such as correlation coefficients and mutual information to identify relevant features. This provided a preliminary selection of features that served as the basis for further refinement.

Subsequently, we utilized wrapper methods, including accounting for feature interactions and refine the feature subset. Wrapper methods involve directly evaluating subsets of features by training the selected machine learning model from the filter selection method. This iterative process allowed us to identify the most impactful features and improve model performance.

After generating multiple models with various feature sets, we compared their performance based on the metrics mentioned above. The evaluation process ensured that the selected model and feature set offered a good balance of achieving the highest TPR and PPV and the lowest FNR and FDR. This approach to combining filter and wrapper methods enabled us to achieve a robust and reliable predictive model, tailored to the specific needs of our application.

III. Results

A. Insight Into Data

The dataset included 200 SCA patients, 10 patients were excluded because they were diagnosed with Thalassemia. The low-risk group included 90 patients while the high-risk group included 110 patients.

B. Preprocessing Results

The analysis of the "ContWithMissing" sheet revealed varying degrees of missing data across different clinical features. To address this, linear interpolation was applied to impute missing values for numerical features using SPSS 21. As shown in Table II, features with more than 10% of missing data were excluded to avoid bias and ensure robust model performance.

TABLE II. Statistics of the Missing Data

	N	Missing	Percent	Minimum	Maximum	Mean
AGE	188	2	1.1%	12	59	31.08
Gender	182	8	4.2%	1	2	1.55
Hb.	190	0	0%	1	13	8.22
WBC	190	0	0%	3.70	45.70	13.97
PIT	190	0	0%	5.10	853	411.81
MCV	189	1	0.5%	37.50	380	86
MCH	189	1	0.5%	19	80	29.128
MCHC	190	0	0%	24.60	38.80	33.64
Ret.	164	26	13.7%	0.014	16.9	2.02
RDW	159	31	16.3%	6.50	208	23.96
Hb A	52	138	72.6%	0	54.10	8.25
Hb A2	103	87	45.8%	1.80	9.40	3.80

Hb S	101	89	46.8%	0	36	8.59
Hb F	127	63	33.2%	0.6	96.1	77.02
AST	141	49	25.8%	13	230	54.18
ALT	147	43	22.6%	14	244	46.28
ALKP	138	52	27.4%	5	352	96.28
GGT	65	125	65.8%	6	223	41.43
Complication	190	0	0%	0	1	0.47

TABLE III. REFORMATE THE FEATURES USING MEDICAL KNOWLEDGE RESULTS

Feature	Term	Range	Value	Define	No. of Missing Estimated
Age (years)	Old Young	12-59	>40 <=40	0 1	2 0
Gender	Male Female		1 2	0 1	0 1
Hb (g/dL.)	Normal Low Very low	2-13	12-15 >=7, <=12 <7	2 1 0	0
WBC (K/uL)	High Normal Low	3.7-45.7	>11.5 >=4.5,<=11.5 <4.5	2 1 0	0
PLT (K/uL)	High Normal Low	5.10-853	>=450 150-450 <=150	2 1 0	0
MCV (fL)	High Normal Low	26.98-380	>94 80>=,<=94 <80	2 1 0	0 1 0
MCH (pg)	High Normal Low	19-80	>31 27>=,<=31 <27	2 1 0	0 1 0
MCHC (%)	High Normal Low	24.6-38.8	>36 32>=,<=36 <32	2 1 0	0
Reticulocyte (%)	High Normal Low	0.014-169	>1.5 0.5>=,<=1.5 <0.5	2 1 0	4 0 22
Red cell distribution width (Male)	High Normal Low	0.01-16.9	>14.5 11.8<=,>=14.5 <11.8	2 1 0	10 0 0
Red cell distribution width (Female)	High Normal Very low	0.08-0.23	>16.1 12.2-16.1 <12	2 1 0	0 0 16
electrophoresis Hb A (%)	Normal Low Very low	0-54.10	95-98 <95 =0	2 1 0	0 77 61
Hb A2 (%)	High Normal	1.8-9.4	>=3.5 1.5-3.5	1 0	27 8
Hb F (%)	Abnormal Normal	0-36	>=2 <2	1 0	89 0
Hb S (%)	Abnormal Normal		1 0	1 0	63 0
Aspartate Amino Transferase (AST) (U/L)	High Normal Low	13-230	>37 15>=,<=37 <15	2 1 0	40 8 1
Alanine Amino Transferase (ALTI) (U/L)	High Normal	14-244	>78 12>=,<=78	2 1	1 42
Alkaline phosphatase (ALPI) (U/L)	High Normal Low	5-352	>150 40>=,<=150 <40	2 1 0	13 36 3
Gamma-glutamyl Transferase GGT (U/L)	High Normal Low	6-223	>85 5,<=85	2 1	25 100
Complications	None Complications		None Complications	0 1	

Age and gender data were nearly complete, with only 1.1% and 4.21% missing, respectively. Critical parameters such as hemoglobin levels, white blood cell count, platelets, mean corpuscular volume, mean cell hemoglobin, and mean cell hemoglobin concentration had no missing values. Reticulocyte and red cell distribution width had 13.68% and 16.32% missing values, respectively, and were thus excluded from the analysis. Hemoglobin electrophoresis tests showed substantial missing data (HbA 72.63%, HbA2 45.79%, HbF 46.84%, HbS 33.16%), leading to their exclusion. Similarly, liver enzyme measurements such as AST, ALT, ALKP, and GGT had significant missing data, particularly GGT with 65.79%, and were excluded from further analysis. This approach ensures that the dataset remains robust and reliable for subsequent analysis and model development. The data was also reformatted using medical knowledge and standard blood count boundaries to categorize values meaningfully. Table III shows the reformat of numerical data using medical knowledge.

C. Filter Selection Results

The feature selection process utilized different methods for the "ContWithMissing" and "CatWithMissing" sheets. For the "ContWithMissing" sheet, Table IV shows the top correlated features using MIC analysis. The top features included Platelets (MIC: 0.41987), WBC (MIC: 0.41139), MCH (MIC: 0.3613), MCV (MIC: 0.3171), MCHC (MIC: 0.27546), Hb (MIC: 0.26462), and AGE (MIC: 0.16802), all of which were included in the final model due to their strong correlations.

For the "CatWithMissing" sheet, the Chi-squared test was used to determine the statistical significance of the features. This analysis revealed several significant features, including WBC and Platelets.

Notably, WBC showed a significant difference between low and high-risk patient categories (p-value: 0.005), and Platelet counts also displayed statistical significance (p-value: 0.027). Table V shows significant features.

By integrating MIC analysis for continuous data and Chi-squared analysis for categorical data, the final model incorporated the most predictive and statistically significant features, thereby enhancing its performance and reliability in predicting mismatch repair complications.

D. Machine Learning Methods Results

The feature selection and ML model evaluation was conducted using the "CatNoMissing" sheet. For this process, 10% of the data was held out as testing data while 90% was used for training. The final dataset underwent ten-fold cross-validation. The LR, SVM, and DT models were assessed using several performance metrics, including TPR, FNR, PPV, and FDR. Table VI shows the final results using feature selection and ML methods.

TABLE IV. TOP CORRELATED FEATURES

Features	MIC (strength)
Platelets	0.41987
WBC	0.41139
MCH	0.3613
MCV	0.3171
MCHC	0.27546
Hb	0.26462
AGE	0.16802

TABLE V. (A) SIGNIFICANT FEATURES

Feature	Medical Knowledge	Patients Categories				Chi-Square Test		
		Low risk		*High risk*		*Frequency*	*Percent*	*p-value*
Age (years)	>40 <=40	79 19	80.6% 10.1%	79 11	87.8% 5.9%	158 30	84% 16%	0.23
Sex	1 2	47 46	50.5% 49.5%	35 54	39.3% 60.7%	82 100	45.1% 54.9%	0.13
Hb (g/dL.)	12-15 >=7, <=12 <7	2 86 12	2% 86% 12%	0 75 15	0% 83.3% 16.7%	2 161 27	1.1% 84.7% 14.2%	0.33
WBC (K/uL)	>11.5 >=4.5,<=11.5 <4.5	49 50 1	49% 50% 1%	63 27 0	70% 30% 0%	112 77 1	35.8% 61.6% 2.6%	**0.005**
Platelet (K/uL)	>=450 150-450 <=150	30 65 5	30% 65% 5%	38 52 0	42.2% 57.8% 0%	68 117 5	35.8% 61.6% 2.6%	**0.027**
MCV (fL)	>94 80>=,<=94 <80	16 47 37	16% 47% 37%	18 49 22	20% 55.1% 24.7%	34 96 59	18% 50.8% 31.2%	0.19
MCH (pg)	>31 27>=,<=31 <27	23 43 34	23% 43% 34%	32 33 24	36% 37.1% 27%	55 76 58	29.1% 40.2% 30.7%	0.15
MCHC (%)	>36 32>=,<=36 <32	6 85 9	6% 85% 9%	7 69 14	7.8% 76.7% 15.6%	13 154 23	6.8% 81.1% 12.1%	0.31
Ret (%)	>1.5 0.5>=,<=1.5 <0.5	5 1 75	6.2% 1.2% 92.6%	6 1 76	7.2% 1.2% 91.6%	11 2 151	6.7% 1.2% 92.1%	1.00

TABLE V (B). SIGNIFICANT FEATURES (CONTINUE)

Feature	Medical Knowledge	Patients Categories				Chi-Square Test		
		Low risk		High risk		Frequency	Percent	p-value
RDW	Male >14.5 11.8<=,>=14.5 <11.8 Female >16.1 12.2<=,>=16.1 <12	84 1 1	97.7% 1.2% 1.2%	72 0 1	98.6% 0 1.4%	156 1 2	98.1% 0.6% 1.3%	1.00
Hb A (%)	95-98 <95 =0	0 9 20	0% 31% 69%	0 11 12	0% 47.8% 52.2%	0 20 32	0 38.5% 61.5%	0.26
Hb A2 (%)	>=3.5 1.5-3.5	34 24	58.6% 41.4%	23 22	51.1% 48.9%	57 46	55.3% 44.7%	0.54
Hb F (%)	>=2 <2	53 5	91.4% 8.6%	41 2	95.3% 4.7%	94 7	93.1% 6.9%	0.69
Hb S (%)	1 0	67 0	100% 0%	60 0	100% 0%	127 0	100% 0	-
AST (U/L)	>37 15>=,<=37 <15	53 24 1	67.9% 30.8% 1.3%	42 20 1	66.7% 31.7% 1.6%	95 44 2	67.4% 31.2% 1.4%	1.00
ALTI (U/L)	>78 12>=,<=78	6 75	7.4% 92.6%	2 64	3% 97%	8 139	5.4% 94.6%	0.29
ALPI (U/L)	>150 40>=,<=150 <40	18 31 25	24.3% 41.9% 33.8%	16 26 22	25% 40.6% 34.4%	34 57 47	24.6% 41.3% 34.1%	1.00
GGT (U/L)	>85 5,<=85	4 32	11.1% 88.9%	1 28	3.4% 96.6%	5 60	7.7% 92.3%	0.37
Complications	Low risk High risk (Complications)					100 90	52.6% 47.4%	

TABLE VI. MACHINE LEARNING METHODS RESULTS

Feature Selection	Model	Features	Training Accuracy	Testing Accuracy	Testing TPR	Testing FNR	Testing PPV	Testing FDR
None	LR	AGE, Gender, Hb, Platelets, WBC, MCV, MCH, MCHC, Ret, RWD, Hb A, Hb A2, Hb F, Hb S, AST, ALTI, ALPI, GGT.	65.3	35[a]	64.3	35.7	30	70
	SVM		56.7	62.5	50	50	46.7	53.3
	DT		58.7	50	64.3	35.7	37.5	62.5
Chi-square	LR	Platelets, WBC.	60.7	52.5	64.3	35.7	39.1	60.9
	SVM		62.7	52.5	54.54	45.45	60	40
	DT		61.9	52.5	54.54	45.45	60	40
MIC	LR	Platelets, WBC, MCH, MCV, MCHC, Hb, AGE.	62	47.5a	85.7	14.3	38.7	61.3
	SVM		54	50	85.7	14.3	40	60
	DT		52	52.5	71.4	28.6	40	60
Wrapper	LR	Platelets, WBC, MCV, MCHC	54.38	52.63	20	80	66.67	33.33
	SVM		60.23	57.90	72.72	27.27	61.53	38.46
	DT		66.08	36.82	36.36	63.63	44.45	55.56
a An overfitting model is a data science concept in which a statistical model fits precisely against its training data.* selected model.								

Initially, models were trained without feature selection, using all available features. The LR model achieved a training accuracy of 65.3% but showed significant overfitting with a testing accuracy of 35%. The TPR and PPV were 64.3% and 30%, respectively, while the FNR and FDR were 35.7% and 70%. The SVM model performed better on the testing data, achieving an accuracy of 62.5%, a TPR of 50%, a PPV of 46.7%, an FNR of 50%, and an FDR of 53.3%. The DT model had a testing accuracy of 50%, with a TPR of 64.3%, a PPV of 37.5%, an FNR of 35.7%, and an FDR of 62.5%.

Next, feature selection was applied using significant features (chi-square) identified through statistical tests, specifically focusing on Platelets and WBC. The LR model improved slightly with a testing accuracy of 52.5%, a TPR of 64.3%, a PPV of 39.1%, an FNR of 35.7%, and an FDR of 60.9%. Both the SVM and DT models showed similar improvements, with SVM achieving a testing accuracy of 52.5%, a TPR of 54.54%, a PPV of 60%, an FNR of 45.45%, and an FDR of 40%. The DT model had a testing accuracy of 52.5%, a TPR of 54.54%, a PPV of 60%, an FNR of 45.45%, and an FDR of 40%.

Further evaluation with the top 7 correlated features (MIC) revealed that the SVM model with these features achieved the best performance metrics, with a testing accuracy of 50%, a TPR of 85.7%, a PPV of 40%, an FNR of 14.3%, and an FDR of 60%. The LR model also performed well, with a testing accuracy of 47.5%, a TPR of 85.7%, a PPV of 38.7%, an FNR of 14.3%, and an FDR of 61.3%. The DT model had a testing accuracy of 52.5%, a TPR of 71.4%, a PPV of 40%, an FNR of 28.6%, and an FDR of 60%.

After applying wrapper selection methods, the SVM model with 4 features demonstrated the best performance among the evaluated models, with a testing accuracy of 57.9%, a TPR of 72.72%, a PPV of 61.53%, an FNR of 27.27%, and an FDR of 38.46%. A total of 2^{18} =262,143 models were generated for each technique and assessed. The final model was selected based on its superior performance metrics, specifically the highest TPR and PPV, along with the lowest FNR and FDR. The selected SVM model showed exceptional capability in correctly identifying high-risk patients while minimizing incorrect classifications. This comprehensive evaluation confirmed the robustness and accuracy of the SVM model in predicting patient risk, making it the optimal choice for this analysis.

IV. DISCUSSION

This study aimed to implement ML algorithms to identify high-risk groups among SCA patients using clinical and pathological data from King Abdulaziz University Hospital. The primary objective was to develop predictive models that enhance the accuracy of risk assessments, thereby improving patient management and treatment outcomes. The application of LR, SVM, and DT in this study demonstrated the potential of ML in healthcare, particularly in the context of genetic disorders like SCA.

The findings indicate that the SVM model outperformed LR and DT in identifying high-risk SCD patients. Specifically, the SVM model with the top 7 correlated features achieved the highest TPR and PPV, along with the lowest FNR and FDR. These results suggest that SVM is particularly effective in distinguishing between high-risk and low-risk patients, making it a valuable tool for clinicians. The experimental results echo theoretical expectations: SVM excelled because the dataset (n = 200, 7–17 active predictors) is high-dimensional relative to sample size, a setting in which the max-margin principle reduces over-fitting. LR performed adequately on linearly separable CBC ratios, whereas DT handled non-linear interactions but was vulnerable to small-sample fragmentation.

A crucial element of the study was the feature selection process, which significantly enhanced model performance. Both filter and wrapper methods were employed to identify the most relevant features, thereby reducing model complexity and improving prediction accuracy. The top 7 correlated features (Platelets, WBC, MCH, MCV, MCHC, Hb, and Age) were found to be significant predictors of risk in SCA patients. Among these, Platelets and WBC counts were particularly noteworthy. Platelets, which are essential for blood clotting, and WBCs, which play a key role in immune response, are critical indicators of a patient's health status and potential complications. Indeed, the complete blood count (CBC) evaluates these cells, helping detect various diseases and conditions [32]. A lower platelet to WBC ratio is associated with increased risk of postoperative infectious complications in patients undergoing radical nephrectomy [33]. In essential thrombocythemia patients, platelet counts outside the normal range correlate with an immediate risk of major hemorrhage, while elevated WBC counts are associated with both thrombosis and hemorrhage [34]. A conserved pattern of recovery, defined by co-regulation of WBC and platelet populations, has been identified across various inflammatory. This pattern of recovery, marked by an exponential decline in WBC and a subsequent linear increase in platelet count, could signify a basic model of human physiological response and serve as a means to identify high-risk patients. [35].

One of the key challenges addressed in this study was the handling of missing data. The application of linear interpolation to impute missing values ensured that the dataset remained comprehensive and informative. Excluding features with more than 10% of missing data helped maintain the integrity of the models, avoiding potential biases that could arise from incomplete data.

The decision to use a combination of filter and wrapper methods for feature selection, rather than embedded methods, was justified by the need for flexibility and computational efficiency. While embedded methods integrate feature selection within the model training process, they can be computationally intensive and less adaptable to different algorithms. The approach adopted in this study allowed for a more tailored feature selection process, enhancing the performance of each ML model.

Despite the promising results, this study has some limitations. The sample size of 200 patients, while adequate for initial model development, may not capture the full spectrum of variability in SCD presentations. Future studies with larger datasets and multi-center collaborations are needed to validate the findings and enhance the generalizability of the models.

Using LR, SVM, and DT with small-sized datasets presents both opportunities and challenges. LR is advantageous for small datasets because it is less prone to overfitting due to its simpler linear model structure. It can provide meaningful insights and interpretability, especially when the relationship between the predictors and the outcome is approximately linear. However, its performance may suffer if the underlying data distribution is complex.

The SVM, particularly with linear kernels, are also suitable for small datasets as they maximize the margin between classes, which can lead to better generalization on unseen data. SVMs are robust to overfitting making them effective for datasets with

a small number of samples but many features. However, SVMs can be computationally expensive, and their performance is highly sensitive to the choice of kernel and hyperparameters, which can be challenging to optimize with limited data.

Finally, DTs were intuitive and easy to interpret, making them appeal for small datasets. They could capture complex interactions between features without requiring a large amount of data. However, DTs are prone to overfitting, especially with small datasets, as they tend to create overly complex trees that do not generalize well.

V. Conclusion

In conclusion, this study demonstrates the potential of ML algorithms, particularly SVM, in identifying high-risk SCA patients. The integration of hybrid feature selection techniques and effective handling of missing data contributed to the development of accurate predictive models. These findings underscore the value of ML in healthcare, offering a tool for risk assessment and patient management in SCA. The capability of ML to enhance clinical decision-making, improving patient outcomes in genetic disorders like SCA. The SVM model (TPR = 85.7 %, PPV = 61.5 %) offers a clinically actionable screening tool for SCA wards. Limitations include (i) single-centre data, which may under-represent regional genotype variability, and (ii) retrospective feature availability, potentially biasing model coefficients. Future work will involve a prospective, multi-centre cohort integrating inflammatory biomarkers and whole-genome variants to refine risk predictions and support personalized therapy pathways.

Acknowledgment

The authors appreciate the associate editors' and anonymous reviewers' suggestions and criticism of an earlier version of this manuscript.

Authors' Contributions

Conceptualization, H.B.; Methodology, H.B., H.A., N.A., S.A.and G.A.; Software, H.B., H.A., and N.A.;Validation, H.B., A.B., and G.A.; Investigation, H.B., H.A., A.A., and G.A.; Resources, H.B., H.A., N.A., A.B., H.A., S. B., A.A., S.A., and G.Z.; Data Curation, H.B., S.A., A.A. and G.Z.; Writing Original Draft Preparation, H.B., N. A.; Writing—Review and Editing, H.B., H.A., N. A., A.B., S.A., and G.Z.; Visualization, H.B.; Supervision, H.B.; Project Administration, H.B.; Funding Acquisition, H.B.

Conflicting of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The Saudi data is restricted to only a few authorized users who are permitted to access the database with Clinical Research Committee (CRC) and Research Ethics Committee (REC) permission. Due to ethical restrictions, de-identified datasets are available upon request to (hrbanjar@kau.edu.sa) for authorization.

References

[1] S. Azar and T. E. Wong, "Sickle Cell Disease: A Brief Update.," Med. Clin. North Am., vol. 101, no. 2, pp. 375–393, Mar. 2017, doi: 10.1016/j.mcna.2016.09.009.

[2] Dhiraj K, "Top 4 advantages and disadvantages of Support Vector Machine or SVM.," Medium, 2019. https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107 (accessed Sep. 24, 2021).

[3] D. J. Weatherall, "The inherited diseases of hemoglobin are an emerging global health burden.," Blood, vol. 115, no. 22, pp. 4331–4336, Jun. 2010, doi: 10.1182/blood-2010-01-251348.

[4] M. Angastiniotis and B. Modell, "Global Epidemiology of Hemoglobin Disorders," Ann. N. Y. Acad. Sci., vol. 850, 1998, [Online]. Available: https://api.semanticscholar.org/CorpusID:10048398.

[5] W. Jastaniah, "Epidemiology of sickle cell disease in Saudi Arabia," Ann. Saudi Med., vol. 31, pp. 289–293, 2011, [Online]. Available: https://api.semanticscholar.org/CorpusID:5382156.

[6] F. B. Piel, M. H. Steinberg, and D. C. Rees, "Sickle Cell Disease.," N. Engl. J. Med., vol. 376 16, pp. 1561–1573, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:35879604.

[7] N. H. Alharbi, R. O. Bameer, S. S. Geddan, and H. M. Alharbi, "Recent Advances and Machine Learning Techniques on Sickle Cell Disease," Futur. Comput. Informatics J., 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:233386170.

[8] P. Sebastiani et al., "A network model to predict the risk of death in sickle cell disease.," Blood, vol. 110 7, pp. 2727–2735, 2007, [Online]. Available: https://api.semanticscholar.org/CorpusID:8718624.

[9] C. T. Quinn, "Minireview: Clinical severity in sickle cell disease: the challenges of definition and prognostication," Exp. Biol. Med., vol. 241, pp. 679–688, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:27272544.

[10] X. W. van den Tweel, J. H. Lee, J. Howard, and K. Fijnvandraat, "Measurement of Disease Severity in Patients with Sickle Cell Disease: A Systematic Review.," Blood, vol. 110, p. 2250, 2007, [Online]. Available: https://api.semanticscholar.org/CorpusID:208430316.

[11] N. R. Shah et al., "Severity Classification for Sickle Cell Disease: A RAND/UCLA Modified Delphi Panel," Blood, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:214553481.

[12] A. Mohammed, P. S. B. Podila, R. L. Davis, K. I. Ataga, J. S. Hankins, and R. Kamaleswaran, "Using Machine Learning to Predict Early Onset Acute Organ Failure in Critically Ill Intensive Care Unit Patients With Sickle Cell Disease: Retrospective Study," J. Med. Internet Res., vol. 22, no. 5, pp. e14693–e14693, May 2020, doi: 10.2196/14693.

[13] V. E. Staartjes, J. M. Kernbach, V. Stumpo, C. H. B. van Niftrik, C. Serra, and L. Regli, "Foundations of Feature Selection in Clinical Prediction Modeling.," Acta Neurochir. Suppl., vol. 134, pp. 51–57, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:244872146.

[14] N. Mlambo, W. K. Cheruiyot, and M. W. Kimwele, "A Survey and Comparative Study of Filter and Wrapper Feature Selection Techniques," 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:38791166.

[15] S. S. M. S and A. Narayanan, "Improving Classification Accuracy Using Combined Filter+Wrapper Feature Selection Technique," 2019 IEEE Int. Conf. Electr. Comput. Commun. Technol., pp. 1–6, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:204816141.

[16] M. Mandal, P. K. Singh, M. F. Ijaz, J. Shafi, and R. Sarkar, "A Tri-Stage Wrapper-Filter Feature Selection Framework for Disease Classification," Sensors (Basel)., vol. 21, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:237341764.

[17] A. Patel et al., "Machine Learning Algorithms in Predicting Hospital Readmissions in Sickle Cell Disease," Blood, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:213979893.

[18] O. B. Ayoade et al., "An Ensemble Models for the Prediction of Sickle Cell Disease from Erythrocytes Smears," EAI Endorsed Trans. Pervasive Heal. Technol., vol. 9, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:262100582.

[19] M. Sarker, "Reinventing Wellness: How Machine Learning Transforms Healthcare," J. Artif. Intell. Gen. Sci. ISSN3006-4023, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:268355442.

[20] F. Cismondia, A. S. Fialho, S. M. Vieira, S. R. Retic, J. M. C. Sousab, and S. N. Finkelstein, "Missing data in medical databases: Impute, delete or classify?," Artif. Intell. Med., vol. 58, no. 1, pp. 63–72, 2013, doi: 10.1016/j.artmed.2013.01.003.

[21] G. Huang, "Missing data filling method based on linear interpolation and lightgbm," J. Phys. Conf. Ser., vol. 1754, no. 1, 2021, doi: 10.1088/1742-6596/1754/1/012187.

[22] "The CBC – providing information about your health," Sonora Quest Laboratories. https://www.sonoraquest.com/patient/knowledge-center/understanding-the-complete-blood-count-cbc/ (accessed Sep. 05, 2021).

[23] H. Banjar et al., "Modelling Predictors of Molecular Response to Frontline Imatinib for Patients with Chronic Myeloid Leukaemia," PLoS One, vol. 12, no. 1, p. e0168947, 2017, doi: 10.1371/journal.pone.0168947.

[24] D. N. Reshef et al., "Detecting Novel Associations in Large Data Sets," Science (80-.)., vol. 334, pp. 1518–1524, 2011, doi: 10.1126/science.1205438.

[25] D. Mindrila and P. Balentyne, "The Chi -Square Test : Analyzing," no. 2013, p. 23, 2013.

[26] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, "Introduction to the Logistic Regression Model," 2005, [Online]. Available: https://api.semanticscholar.org/CorpusID:59226812.

[27] H.-Y. Kim, "Statistical notes for clinical researchers: logistic regression," Restor. Dent. \& Endod., vol. 42, pp. 342–348, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:38386217.

[28] "Tutorial on Support Vector Machine (SVM)," 2021. Accessed: Sep. 24, 2021. [Online]. Available: https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf.

[29] D. Boswell, "Introduction to Support Vector Machines," 2002, [Online]. Available: https://api.semanticscholar.org/CorpusID:18986102.

[30] Z. Saurav, M. M. Mitu, N. S. Ritu, M. A. Hasan, S. Arefin, and D. M. Farid, "A New Method for Learning Decision Tree Classifier," 2023 Int. Conf. Electr. Comput. Commun. Eng., pp. 1–6, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:258219459.

[31] P. L. Bokonda, K. Ouazzani-Touhami, and N. Souissi, "Predictive analysis using machine learning: Review of trends and methods," 2020 Int. Symp. Adv. Electr. Commun. Technol., pp. 1–6, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:237376199.

[32] S. Gajbhiye and J. R. Aate, "Blood Report Analysis-A Review," Trop. J. Pharm. Life Sci., 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:264109151.

[33] A. Garbens et al., "Platelet to white blood cell ratio predicts 30-day postoperative infectious complications in patients undergoing radical nephrectomy for renal malignancy.," Can. Urol. Assoc. J., vol. 11 11, pp. E414–E420, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:31726668.

[34] P. J. Campbell et al., "Correlation of blood counts with vascular complications in essential thrombocythemia: analysis of the prospective PT1 cohort.," Blood, vol. 120 7, pp. 1409–1411, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:206911231.

[35] B. H. Foy, T. M. Sundt, J. C. T. Carlson, A. D. Aguirre, and J. M. Higgins, "White Blood Cell and Platelet Dynamics Define Human Inflammatory Recovery," medRxiv, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:235602574.