Machine Learning-Based Disease Severity Prediction in Sickle Cell Patients: Spectroscopic Insights

Sumit Kumar Roy, Saurabh Gupta and Pankaj Jain
Department of Biomedical Engineering, National Institute of Technology Raipur, Raipur, Chhattisgarh, India

Keywords: Sickle Cell Disease, Sickle Cell Anemia, High Performance Liquid Chromatography, Artificial Intelligence, Machine Learning, Spectroscopy.

Abstract: Sickle cell disease (SCD) presents a significant health challenge with diverse clinical manifestations. Early and accurate prediction of the onset and severity of co-morbidities in SCD is vital for improving outcomes. In this study, we employ advanced healthcare informatics, and machine learning techniques to analyze longitudinal blood pathology data. By focusing on crucial hematological parameters, we gain valuable insights into SCD's pathophysiology. Additionally, incorporating spectroscopic insights into the study unveils molecular details, enriching the understanding of the disease's complexity and paving the way for more nuanced and targeted interventions. Utilizing this data, we construct predictive models enabling personalized interventions and advancing precision healthcare management. The research revealed that Random Forest outperforms other algorithms, achieving an accuracy of 88%, recall of 82%, and specificity of 92%. This robust evaluation underscores the model's reliability in predicting both positive and negative instances. These findings offer a promising pathway for enhancing disease prediction, management, and treatment planning, providing invaluable guidance for clinical practice in the context of sickle cell disease.

1 INTRODUCTION

Predicting the onset and severity of co-morbidities in patients with sickle cell disease (SCD) is a paramount challenge, underpinned by the potential to enhance precision medicine and optimize patient outcomes. SCD, an autosomal recessive monogenic disorder characterized by specific mutations in the β-globin gene, engenders the polymerization of abnormal hemoglobin S (HbS) molecules, ultimately inducing sickling in red blood cells (RBCs) (Kato et al., 2018), (Dheyab et al., 2020).

The pathophysiological underpinnings of this disorder entail pronounced hematological perturbations, including alterations in hemoglobin concentration (Hb), reticulocyte count (RC), hematocrit (Hct), and red blood cell indices (Kato et al., 2018), (Dheyab et al., 2020). As SCD affects millions worldwide, the spectrum of clinical manifestations exhibits substantial heterogeneity that is significantly modulated by dynamic variations within these hematological parameters such as iron deficiency over time (Liu et al., 2021), (da Silva et al., 2020) . In the contemporary landscape of advanced healthcare informatics, Spectroscopy, and machine learning (ML), the systematic exploration of longitudinal blood pathology data enables the formulation of quantitative models for predictive analytics, thereby advancing the precision of healthcare for SCD patients (Elsabagh et al., 2023), (Farota et al., 2022).

Hb, is a crucial indicator of anemia severity in SCD patients. Levels below 11.0 g/dL signify anemia. RC, a fraction of circulating RBCs, reflects erythropoietic activity and hemolysis. Hct, representing the percentage of RBCs in total blood volume (38% to 52% norm), provides insights into anemia. Utilizing advanced computational methods, the analysis of hemoglobin dynamics assists in forecasting the onset and intensity of anemia in SCD patients (Kato et al., 2018). By applying these techniques to RC data, predictive models are crafted to assess the risk of acute anemia events (Dheyab et al., 2020), (Elsabagh et al., 2023). Additionally, ML models, which capture Hct trends, provide insights into the trajectory of anemia (Elsabagh et al., 2023).

RBCs indices like mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and mean corpuscular hemoglobin concentration (MCHC) indicate RBC size and hemoglobin content. ML techniques can decode various blood in-

Roy, S., Gupta, S. and Jain, P.
Machine Learning-Based Disease Severity Prediction in Sickle Cell Patients: Spectroscopic Insights.
DOI: 10.5220/0012343800003651

In *Proceedings of the 12th International Conference on Photonics, Optics and Laser Technology (PHOTOPTICS 2024)*, pages 51-56
ISBN: 978-989-758-686-6; ISSN: 2184-4364

dices, offering valuable predictions about the evolving pathology, disease manifestations, and complications in SCD (Elsabagh et al., 2023), (Farota et al., 2022).

Advanced computational tools, including deep neural networks and ensemble methods, offer quantitatively-driven insights into the progression and risks of SCD-related complications based on longitudinal blood pathology data. These models operate by recognizing complex patterns in the temporal evolution of hematological parameters, ultimately enabling the timely and targeted implementation of individualized interventions (Elsabagh et al., 2023), (Farota et al., 2022), (Gollapalli et al., 2022).

In the realm of related research, Farota et al. introduced a predictive model that combines five classification algorithms, including AdaBoost, Logistic Regression, Support Vector Machine (SVM), k-nearest neighbors (KNN), and Random Forest (RF). Their evaluation showcased high accuracy, particularly 1 for SVM, RF, and 0.95 for Logistic Regression. Notably, all classifiers, except K-NN, exhibited an AUC close to 1, emphasizing the robustness of their predictive capabilities (Elsabagh et al., 2023), (Farota et al., 2022). In a similar vein, Gollapalli et al. developed a data-driven machine learning model based on hospital data from clinical SCD patients. They revealed that acute chest syndrome encompasses symptoms such as chest pain, coughing, high fever, hypoxia, and lung infiltrates, often resulting from sickling in the lungs tiny blood vessels, causing pulmonary infarction. Their research indicated that SCD patients with acute chest pain typically require hospitalization for 3 to 14 days (Farota et al., 2022).

Meanwhile, Liu et al. presents a microfluidic-based approach with on-chip gas control for the impedance spectroscopy of suspended cells within the frequency range of 40 Hz to 110 MHz. A comprehensive bioimpedance of sickle cells under both normoxia and hypoxia is achieved rapidly (within 7 min) and is appropriated by small sample volumes (2.5 μL) (Liu et al., 2021). Roy et al. proposed a method using blood smear images to predict the percentage of sickling and establish a form factor of 1.81. Their study utilized data from blood smears of diseased patients obtained from a medical center (Sumit Kumar Roy and Tyagi, 2020).

This work embarks on a quantitative journey into the predictive power of different ML techniques, driven by longitudinal blood pathology data, in anticipating the emergence and severity of diverse diseases or co-morbidities in SCD patients. A multidisciplinary assembly of experts in hematology, data science, and clinical research collaborates to unravel the intricate relationships between these parameters and disease outcomes, aiming to articulate quantitative challenges and propose innovative methodologies. The quantitative insights generated from this research hold the potential to revolutionize the management of SCD, facilitating precise, data-informed healthcare strategies that improve patient outcomes.

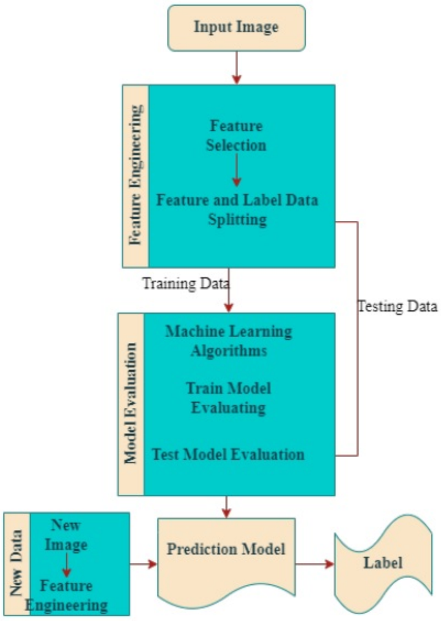

Figure 1: Process flow of the work.

2 MATERIALS AND METHODS

2.1 Methodology

Our approach follows a systematic procedure for classification of severity of diseases in Sickle cell disease (SCD) patients, commencing with data collection and meticulous pre-processing to guarantee data quality. Subsequently, we engage in feature engineering to select the most informative and pertinent features, thereby enabling the machine learning model to learn effectively and yield precise predictions. Following this, we partition the data into training and testing sets to facilitate the training and evaluation of the selected machine learning algorithms. If the algorithms' performance falls short of acceptable criteria, they undergo fine-tuning and retraining until they meet the desired performance benchmarks. Finally, the trained algorithms are deployed and used to predict outcomes on new data (Elsabagh et al., 2023), (Farota et al., 2022). The process flow is illustrated in Figure 1 (Jain and Gupta, 2023).

2.2 Dataset Preparation

The research employed a SCD dataset sourced from the Sickle Cell Institute Chhattisgarh, Raipur, comprising information from 63 Sickle Cell patients. In order to evaluate the ML models' performance in predicting the severity of SCD in patients, the dataset includes a range of parameters such as patient names, ages, appointment dates, and a range of hematological indicators: Hemoglobin (HGB), Red Blood Cells (RBC), Hematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Red Cell Distribution Width-CV (RDW-CV), Red Cell Distribution Width-SD (RDW-SD), White Blood Cells (WBC), Neutrophil Count (Neut#), Lymphocyte Count (Ly#), Mid-cell Count (Mid#), Neutrophil Percentage (Neut%), Lymphocyte Percentage (Ly%), Mid-cell Percentage (Mid%), Platelet Count (PLT), Mean Platelet Volume (MPV), Platelet Distribution Width (PDW), Platelet Crit (PCT), and Platelet Large Cell Ratio (PLCR) (Dheyab et al., 2020), (Liu et al., 2021). The research also gathered absorbance spectra data from absorption spectroscopy encompassing positive samples as well as negative samples within the wavelength range spanning from 395 nm to 750 nm. Subsequent refinement of the data involved concentrating on absorbance spectra specifically between 400 nm and 600 nm, ultimately forming the dataset (Srivastava et al., 2021).

2.3 Feature Engineering

In the study, various parameters from the SCD dataset were employed as features to extract meaningful information, and enhance the effectiveness of ML algorithms by accurately representing the data. Utilizing a variety of features bolstered the outcomes and efficiency of our ML system. This method not only streamlines computational processes but also enriches the interpretability of ML algorithms by encompassing diverse facets of sickle cell data. Moreover, it guards against overfitting, enhancing the accuracy and reliability of our models and resulting in more dependable and broadly applicable predictions (Das et al., 2019).

To ascertain the most crucial features, the study incorporated six essential parameters including HGB, RBC, HCT, MCV, WBC, and PLT. For the remaining features, the ANOVA method was applied, allowing for both dimensionality reduction and the selection of the most informative features. These chosen features were then divided into two datasets using K-fold cross-validation with a value of K=5. This cross-validation technique guaranteed rigorous evaluation by iteratively partitioning the data into training and testing sets (Elsabagh et al., 2023), (Das et al., 2019).

The training dataset was employed to instruct the ML model, allowing it to grasp the underlying data patterns, while the independent testing dataset was used to assess the model's performance and gauge its predictive accuracy. This meticulous approach guaranteed a dependable evaluation of the ML model's efficacy in predicting outcomes (Das et al., 2019), (Wahed et al., 2022).

2.4 Machine Learning Algorithms

Machine learning has transformed numerous fields by harnessing its ability to autonomously learn from historical data and make predictions, eliminating the need for explicit programming. Its integration into medical applications has led to significant advancements in disease diagnosis, treatment planning, and drug development. Machine learning algorithms are broadly categorized into supervised learning and unsupervised learning (Khalaf et al., 2017), (Das et al., 2019), (Wahed et al., 2022).

In this work, study focused on 5 different SL algorithms, namely: Random Forest (RF), Linear Support Vector Machine (LSVM), Radial Support Vector Machine (RSVM), Polynomial Support Vector Machine (PSVM), and Sigmoid Support Vector Machine (SSVM) (Wahed et al., 2022), (Saturi, 2023). The use of these algorithms provides a structured framework for predicting disease severity or co-morbidities in SCP, utilizing labeled data to train and evaluate the models.

Support vector machines serve as robust supervised learning tools, excelling in classification, regression, and outlier detection. They thrive in high-dimensional spaces, offer adaptability with diverse kernels, and optimize memory usage through support vectors (Liu et al., 2021), (Wahed et al., 2022). A Decision Tree resembles a tree-shaped flowchart, dividing training data into smaller subgroups, with each subgroup marked by a class label on the terminal node. Random Forest constructs decision trees based on input data, evaluates multiple trees, and selects the best solution through voting. This method enables us to fully leverage the potential of ML, offering accurate and reliable predictions for disease severity. These outcomes contribute significantly to the progress in medical research and clinical practice.

Table 1: Evaluation metrics of our mode.

Metrices	Methods
Accuracy	$\frac{tsP+tsN}{tsP+tsN+fsP+fsN}$
Recall	$\frac{tsP}{tsP+fsN}$
Precision	$\frac{tsP}{tsP+fsP}$
Specificity	$\frac{tsN}{tsN+fsP}$
Classification Error or Error sample rate	1- Accuracy

2.5 Performance Metrics

Table 1 presents the evaluation metrics (accuracy, recall, precision, specificity, and classification error) employed to assess our predictive models. These metrics are exclusively evaluated on the entire test dataset. The parameters tsP, tsN, fsP, and fsN represent true positive, true negative, false positive, and false negative, derived from the confusion matrix (Jain and Gupta, 2023).

3 RESULTS AND DISCUSSION

In this study, data from 63 Sickle cell disease (SCD) patients was categorized into three classes of disease severity (0 = not severe, 1 = severe and 2 = high severe). The effectiveness of each machine learning model was assessed through a series of experiments aimed at predicting disease severity in these patients.

The classification process encompasses four key phases as shown in figure1. In the initial data preparation phase, a comprehensive dataset was created, including all features previously discussed. Subsequently, feature selection was applied to this dataset, incorporating six essential parameters, while the remaining features were evaluated using ANOVA. This process resulted in the selection of the 12 most relevant features, forming a new feature dataset.

Machine learning algorithms were then trained using this new feature dataset to build a disease detection classifier. The study utilized Random Forest, Linear Support Vector Machine, Radial Support Vector Machine, Polynomial Support Vector Machine, and Sigmoid Support Vector Machine. The effectiveness of each predictive machine learning model for disease severity was evaluated within relevant contexts.

Using 5-Fold Cross-Validation, which systematically divides the feature dataset into five non-overlapping subsets, the classifiers were assessed. A classifier's efficiency heavily relies on data quality, with its own set of advantages and limitations. Performance measurements for machine learning efficiency in this study encompassed the confusion matrix, accuracy, recall, precision, and specificity. Additionally, classification error was calculated for predictive performance.

Table 2: Comparison of performance parameter.

Classifiers	Accuracy	Recall	Classification error
RSVM	0.817	0.727	0.183
LSVM	0.742	0.586	0.258
SSVM	0.729	0.600	0.271
PSVM	0.667	0.500	0.333

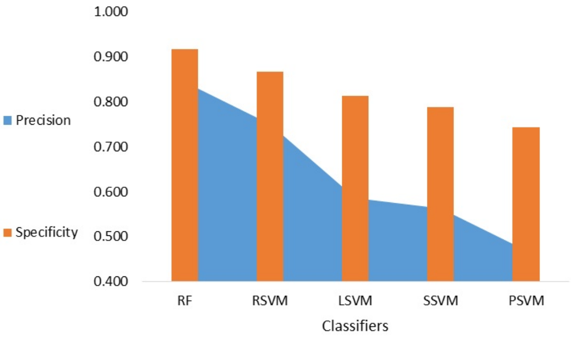

Figure 2: Comparison performance of the classification model using specificity and precision.

Table 2 presents a comparison of accuracy, recall, and classification error as performance measures for different classifiers. Furthermore, a performance evaluation of the classification models was conducted based on specificity and precision, as illustrated in Figure 2.

According to this study on multi-labeled SCD dataset, RF results outperformed those from other SVM classifiers, due to its aptitude for managing intricate, high-dimensional datasets featuring noisy or irregular patterns. In Figure 2 and Table 1, five classification algorithms' performance is assessed using various metrics. Elevated accuracy signifies strong model performance, while high precision and recall values indicate the model's appropriateness for a particular task. High specificity values indicate the model's proficiency in recognizing negative instances. Both RF and RSVM achieved good accuracy (88.2 and 81.7%), precision (84 and 75%), and recall (82 and 73%), showing their reliability in predicting both positive and negative instances. RSVM specificity (87%) is relatively low compared to RF (92%) because RSVM is less effective at predicting negative instances. Furthermore, RF classification error (12%) is relatively lower than RSVM (18%), suggesting that the model makes fewer mistakes in predicting outcomes than RSVM.

RF and RSVM are high-performing classifiers, with RF outperforming RSVM in accuracy, specificity, and classification error. All this because, RF employs ensemble learning, uniting numerous decision trees to effectively capture intricate SCD data relationships while minimizing over-fitting. It stands out in handling noisy data and outliers, making it robust in real-world situations where data quality is a concern. RF also provides valuable insights into feature importance, aiding in the identification of pivotal variables. Additionally, its capacity for parallelization ensures the efficient processing of sizable datasets. In contrast, although R-SVM can handle non-linear relationships using the radial basis function, it may necessitate meticulous parameter tuning and feature scaling, rendering it somewhat more intricate in specific contexts.

Also RSVM tends to outperform LSVM, PSVM, and SSVM models due to its adeptness at managing intricate, non-linear data relationships. While LSVM is confined to straight lines or hyper planes for class separation, RSVM employs the radial basis function kernel, enabling it to transform data into a higher-dimensional space. In this space, complex non-linear relationships are more accurately captured. Although PSVM and SSVM also employ non-linear kernels, they often struggle with intricate data patterns. RSVM, with its radial basis function kernel, excels in scenarios where class boundaries are intricate and not easily defined geometrically. Its ability to adapt to data intricacies results in a more flexible and accurate decision boundary.

While previous studies have made significant strides in SCD and ML, our research introduces novelty by utilizing real-time blood pathology SCD data and diverse ML techniques. This approach is crucial for disease severity/co-morbidities prediction, aiding in diagnosis, disease monitoring, drug development, regenerative medicine, and fundamental research. The findings presented in the study also open avenues for future research in the field of inherited blood disorders. One potential direction involves the exploration of advanced spectroscopic methods, with a focus on refining techniques for real-time monitoring and diagnosis which might also consider the integration of multi-omics approaches, combining spectroscopic insights with genomics, transcriptomics, and metabolomics data to provide a more comprehensive understanding of the molecular intricacies underlying these disorders.

4 CONCLUSION

This study conducted a comparative analysis of five distinct machine learning techniques: Random Forest, Linear Support Vector Machine, Radial Support Vector Machine, Polynomial Support Vector Machine, and Sigmoid Support Vector Machine for classifying disease severity in sickle cell patients. The system predicts disease severity, guiding treatment and medication dosage. Performance metrics were assessed across all classifiers, revealing Random Forest as the most accurate method with 88% accuracy, 82% recall, and 92% specificity. The study's stability and reliability were affirmed through performance evaluation. Future work may explore more features from advanced spectroscopic methods and also deep learning techniques for classification, contingent on obtaining sufficient training data to harness deep learning's full potential.

ACKNOWLEDGEMENTS

We wish to thank the Sickle cell Institute (SCIC) Raipur, Chhattisgarh, for providing pathology hematological data to conduct this research work.

REFERENCES

da Silva, W. R., Silveira Jr, L., and Fernandes, A. B. (2020). Diagnosing sickle cell disease and iron deficiency anemia in human blood by raman spectroscopy. *Lasers in Medical Science*, 35(5):1065–1074.

Das, P. K., Meher, S., Panda, R., and Abraham, A. (2019). A review of automated methods for the detection of sickle cell disease. *IEEE reviews in biomedical engineering*, 13:309–324.

Dheyab, H. F., Ucan, O. N., Khalaf, M., and Mohammed, A. H. (2020). Implementation a various types of machine learning approaches for biomedical datasets based on sickle cell disorder. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6. IEEE.

Elsabagh, A., Elhadary, M., Elsayed, B., Elshoeibi, A. M., Ferih, K., Kaddoura, R., Alkindi, S., Alshurafa, A., Alrasheed, M., Alzayed, A., et al. (2023). Artificial intelligence in sickle disease. *Blood Reviews*, page 101102.

Farota, S. B., Diallo, A. H., Ba, M. L., Camara, G., and Diagne, I. (2022). An ai-based model for the prediction of a newborn's sickle cell disease status. In *International Conference on Innovations and Interdisciplinary Solutions for Underserved Areas*, pages 96–104. Springer.

Gollapalli, M., Alabdullatif, L., Alsuwayeh, F., Aljouali, M., Alhunief, A., and Batook, Z. (2022). Text mining on hospital stay durations and management of sickle cell disease patients. In *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 1–6. IEEE.

Jain, P. and Gupta, S. (2023). Multi-exposure laser speckle contrast imaging (meci)-based prediction of blood flow using random forest (rf) with k-means (km). *Cureus*, 15(6).

Kato, G. J., Piel, F. B., Reid, C. D., Gaston, M. H., Ohene-Frempong, K., Krishnamurti, L., Smith, W. R., Panepinto, J. A., Weatherall, D. J., Costa, F. F., et al. (2018). Sickle cell disease. *Nature reviews Disease primers*, 4(1):1–22.

Khalaf, M., Hussain, A. J., Keight, R., Al-Jumeily, D., Fergus, P., Keenan, R., and Tso, P. (2017). Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. *Neurocomputing*, 228:154–164.

Liu, J., Qiang, Y., and Du, E. (2021). Dielectric spectroscopy of red blood cells in sickle cell disease. *Electrophoresis*, 42(5):667–675.

Saturi, S. (2023). Review on machine learning techniques for medical data classification and disease diagnosis. *Regenerative Engineering and Translational Medicine*, 9(2):141–164.

Srivastava, S., Srinivasan, R., Nambison, N. K., Gorthi, S. S., et al. (2021). Diagnosis of sickle cell anemia using automl on uv-vis absorbance spectroscopy data. *arXiv preprint arXiv:2111.12711*.

Sumit Kumar Roy, Hemlata Sinha, P. M. and Tyagi, D. (2020). Poikilocyte cell detection in microscopic images of blood smears using image processing techniques. *International Journal on Emerging Technologies*, 11(3):234–239.

Wahed, F. F., Juliette A, A., Sinthia, P., and Mary, G. (2022). Detection of sickle cell anemia using svm classifier. In *AIP Conference Proceedings*, volume 2405. AIP Publishing.