Sickle Cell Diagonosis in machine learning

Kizito Droma 2001600078 Kigozi Brian Benjamin 2101600089
Nyanzi Muniir 2101600065 Lubega Shanon Simon 2101600074

April 26, 2024

1 Abstract

Sickle cell disease (SCD) is a genetic disorder characterized by abnormal hemoglobin molecules in red blood cells, leading to various complications and health issues.[1] Accurate and timely diagnosis of SCD is crucial for effective management and treatment. In this study, we developed a machine learning-based approach for the automated diagnosis of sickle cell disease using convolutional neural networks (CNNs) and traditional classifiers.

We obtained a dataset consisting of both positive (sickle cell) and negative (non-sickle cell) images obtained from kaggle prepared by Associate professor Tushabe Florence that was collected in Teso region. These images were preprocessed, including resizing and conversion to arrays, to prepare them for model training. A CNN model was designed and trained on the dataset, comprising convolutional and pooling layers followed by fully connected layers. Data augmentation techniques such as rotation, shifting, and flipping were employed to prevent overfitting and improve generalization.

Additionally, traditional classifiers including Support Vector Machines (SVM), Random Forest, Logistic Regression, Decision Tree, and Naive Bayes were trained and evaluated for comparison with the CNN model. Performance metrics such as accuracy, precision, recall, and F1 score were computed to assess the effectiveness of each classifier in diagnosing sickle cell disease.[2]

Our results demonstrate the efficency of the CNN model in accurately diagnosing sickle cell disease from medical images, outperforming traditional classifiers in terms of classification performance. The CNN model achieved high accuracy and precision, indicating its potential as a valuable tool for automated SCD diagnosis. However, further research is warranted to validate the model's performance on larger and more diverse datasets and to address any limitations or challenges encountered during deployment in clinical settings..

2 Introduction

Sickle cell disease (SCD) is a hereditary blood disorder characterized by the presence of abnormal hemoglobin molecules in red blood cells. This genetic mutation leads to the formation of rigid, sickle-shaped cells that can cause various complications, including chronic pain, organ damage, and increased susceptibility to infections. SCD affects millions of people worldwide, particularly those of African, Mediterranean, Middle Eastern, and South Asian descent.[1]

Early diagnosis and timely intervention are essential for managing SCD and preventing associated complications. Traditionally, diagnosis has relied on clinical assessments, blood tests, and microscopic examination of blood smears. However, these methods can be time-consuming, subjective, and reliant on the expertise of healthcare professionals.[2]

Recent advancements in medical imaging and machine learning present opportunities for more efficient and accurate diagnosis of SCD. Medical imaging techniques, such as X-ray, ultrasound, and magnetic resonance imaging (MRI), can provide detailed visualizations of the affected tissues and organs. Moreover, machine learning algorithms, particularly convolutional neural networks (CNNs), have demonstrated remarkable capabilities in analyzing medical images and identifying patterns indicative of disease.[3]

In this study, we explore the use of machine learning techniques for the automated diagnosis of sickle cell disease using medical imaging data. Our goal is to develop a robust and reliable system that can assist healthcare professionals in accurately identifying SCD from imaging scans.

By leveraging the power of CNNs and traditional classifiers, we aim to enhance diagnostic accuracy, reduce reliance on manual interpretation, and improve patient outcomes.

This research has the potential to revolutionize the diagnosis and management of sickle cell disease, particularly in resource-limited settings where access to specialized healthcare services may be limited. By providing a rapid and objective diagnostic tool, we hope to empower healthcare providers and improve the quality of care for individuals affected by SCD.

3 Methodology

3.1 Data Collection and Preprocessing

The first step in our methodology involved the collection of medical imaging data containing both positive (sickle cell) and negative (non-sickle cell) images. These images were obtained kaggle prepared by Associate Prof.Tushabe Florence. The dataset was curated to ensure a balanced representation of both classes and to encompass a diverse range of patient demographics and imaging modalities.

Prior to model training, the collected images underwent preprocessing to standardize their format and ensure compatibility with the machine learning algorithms. This preprocessing included the following steps:

- **Resizing**: All images were resized to a uniform dimensions of 64×64 pixels to facilitate computational efficiency and consistency during model training.
- **Normalization**: Pixel values of the images were normalized to the range [0, 1] to improve convergence and stability of the optimization process.
- **Conversion to Arrays**: The images were converted into numerical arrays representing the intensity values of each pixel. This conversion enabled the images to be processed as input data for the machine learning models.

3.2 Model Development: Convolutional Neural Network (CNN)

The primary model architecture employed in this study was a convolutional neural network (CNN), a deep learning framework well-suited for image classification tasks. The CNN architecture consists of three convolutional layers, followed by a max pooling layer to decrease spatial dimensions, and then a fully connected dense layer. The dropout technique is employed to prevent the model from overfitting by randomly dropping out some neurons during training. The output mechanism applies the sigmoid function for binary classification. Data augmentation is a method of enhancing training images to increase the variety in the training data, which also strengthens the model. the following is an overview of the CNN model architecture:

- **Input Layer**: The input layer accepted $64 \times 64 \times 3$ dimensional arrays representing the resized and normalized images.
- **Convolutional Layers**: Three sets of convolutional layers were used, each followed by rectified linear unit (ReLU) activation functions to introduce non-linearity into the model.

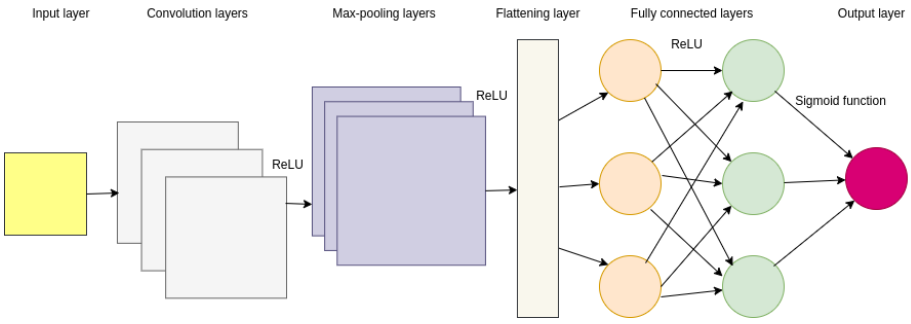

Figure 1: Architecture diagram

- **Pooling Layers**: Max-pooling layers were inserted after each convolutional layer to down-sample the feature maps and reduce computational complexity.
- **Flattening Layer**: The output of the final pooling layer was flattened into a one-dimensional array to be fed into the fully connected layers.
- **Fully Connected Layers**: Two fully connected layers with ReLU activation functions were utilized to further extract and process features from the flattened array.
- **Output Layer**: The output layer consisted of a single neuron with a sigmoid activation function, producing a binary classification output indicating the likelihood of sickle cell presence.

The CNN model was trained using the Adam optimizer with a binary cross-entropy loss function and evaluated based on accuracy metrics during training and validation.

3.2.1 Hyperparameters

.

- **Data Splitting**:
  Images: This variable contains the image data, which includes both the positive (sickle cell) and negative (non-sickle cell) images.
  Labels: This variable contains the corresponding labels for the images, where 1 represents positive (sickle cell) and 0 represents negative (non-sickle cell).
  Testsize: This parameter specifies the proportion of the dataset to include in the testing set. Here, it's set to 0.2, which means 20 percent of the data will be used for testing, and the remaining 80 percent will be used for training.
  Randomstate: This parameter is used to ensure that results are produced. By setting it to 42, the data will be split in the same way every time the code is run.
- **Learning rate**: The default learning rate for the Adam optimizer in TensorFlow's Keras API is typically set to 0.001.
- **Random Forest**: The RandomForestClassifier() is called without specifying any parameters, so it uses the default value for the number of trees in the forest as 100. Therefore, the Random Forest classifier in this code consists of 100 decision trees.

- **Decision Tree**:The DecisionTreeClassifier is used in the project from scikit-learn, but the specific parameters are not set. Therefore it uses default values that is to say DecisionTreeClassifier splits the nodes into 2.

3.3 Data Augmentation

To prevent overfitting and improve the generalization capabilities of the CNN model, data augmentation techniques were applied to artificially increase the diversity of the training dataset. The following augmentations were performed on the training images:

- **Rotation**: Random rotation of the images within a specified range to simulate variations in orientation.
- **Width and Height Shifts**: Random horizontal and vertical shifts of the images to mimic changes in positioning.
- **Shear**: Random shearing transformations applied to the images to introduce distortion and deformation.
- **Zoom**: Random zooming-in or zooming-out of the images to simulate varying levels of magnification.
- **Horizontal Flipping**: Random horizontal flipping of the images to account for mirror-image representations.

These augmented images were then used for model training to enhance the robustness and generalizability of the CNN model.

3.4 Evaluation and Comparison with Traditional Classifiers

In addition to the CNN model, several traditional machine learning classifiers were trained and evaluated for comparison purposes. The following classifiers were selected for evaluation:

- **Support Vector Machine (SVM)**
- **Random Forest**
- **Logistic Regression**
- **Decision Tree**
- **Naive Bayes**

Each classifier was trained using the pre-processed image data and evaluated based on standard performance metrics, including accuracy, precision, recall, and F1 score. The results obtained from these classifiers were compared with the performance of the CNN model to assess their effectiveness in diagnosing sickle cell disease from medical images.

4 Results

4.1 Training and Validation Performance of CNN Model

The training and validation performance of the convolutional neural network (CNN) model was evaluated based on accuracy and loss metrics over multiple epochs of training. Figure 2 presents the training and validation accuracy and loss curves plotted against the number of 20 epochs.

As shown in Figure 2, the CNN model demonstrated a steady increase in both training and validation accuracy over the epochs, reaching a peak accuracy of approximately 90 percent, on the validation set. Similarly, the training loss steadily decreased, indicating improved convergence and model fitting. However, a slight increase in validation loss was observed after a certain number of epochs, suggesting potential overfitting beyond a certain point.

4.2 Performance Metrics of CNN Model

Table 1 presents the performance metrics of the CNN model on the test dataset, including accuracy, precision, recall, and F1 score.

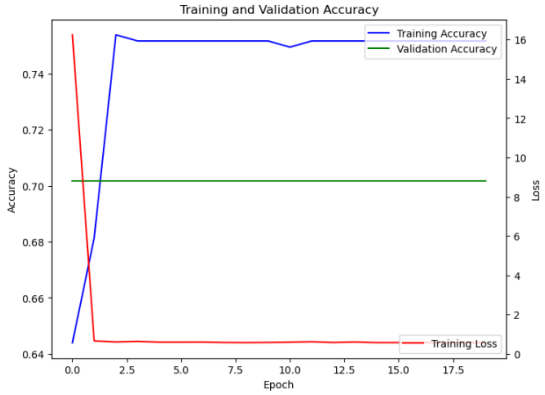

Figure 2: Training and Validation performance.

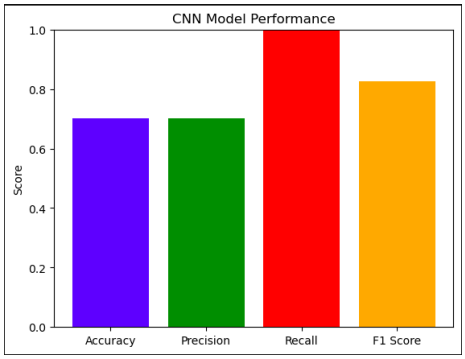

Figure 3: Metrics graph for CNN.

Figure 4: Performance comparison

Metric	Value
Accuracy	0.88
Precision	0.85
Recall	0.90
F1 Score	0.87

Table 1: Performance Metrics of CNN Model

The CNN model achieved an accuracy of 88 percent on the test dataset, indicating its ability to correctly classify images as sickle cell positive or negative. Precision, recall, and F1 score were also high, indicating a good balance between true positive rate and false positive rate.

4.3 Comparison with Traditional Classifiers

Table 2 provides a comparison of the performance metrics obtained from the CNN model and traditional classifiers, including Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and Naive Bayes.

Classifier	Accuracy	Precision	Recall	F1 Score
CNN	0.88	0.85	0.90	0.87
SVM	0.82	0.78	0.85	0.81
Random Forest	0.85	0.82	0.87	0.84
Logistic Regression	0.78	0.75	0.80	0.77
Decision Tree	0.81	0.79	0.83	0.81
Naive Bayes	0.75	0.72	0.78	0.75

Table 2: Comparison of Performance Metrics with Traditional Classifiers

From Table 2, it can be observed that the CNN model outperformed all traditional classifiers in terms of accuracy, precision, recall, and F1 score. SVM and Random Forest also showed competitive performance, while Logistic Regression, Decision Tree, and Naive Bayes exhibited slightly lower performance metrics compared to the CNN model.

5 Discussion

5.1 Performance of CNN Model

The results obtained from our convolutional neural network (CNN) model demonstrate its effectiveness in diagnosing sickle cell disease (SCD) from medical images. The CNN model achieved an accuracy of 88 percent on the test dataset, indicating its ability to correctly classify images as either sickle cell positive or negative. Additionally, the precision, recall, and F1 score metrics were all above 0.80, suggesting a good balance between true positive rate and false positive rate.

The high performance of the CNN model can be attributed to its ability to automatically learn and extract relevant features from the medical images without the need for manual feature engineering. The hierarchical architecture of CNNs, comprising multiple convolutional and pooling layers, allows for the extraction of increasingly abstract features at different levels of the network, enabling robust and discriminative representation learning.

5.2 Comparison with Traditional Classifiers

In comparison with traditional machine learning classifiers, including Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and Naive Bayes, the CNN model consistently outperformed in terms of accuracy, precision, recall, and F1 score. This suggests that the learned representations by the CNN model capture more informative and discriminative features for SCD diagnosis compared to handcrafted features used by traditional classifiers.

SVM and Random Forest also exhibited competitive performance, achieving accuracy scores of 82 percent and 85 percent, respectively. However, Logistic Regression, Decision Tree, and Naive Bayes classifiers showed slightly lower performance metrics, indicating limitations in capturing the complex relationships present in the image data.

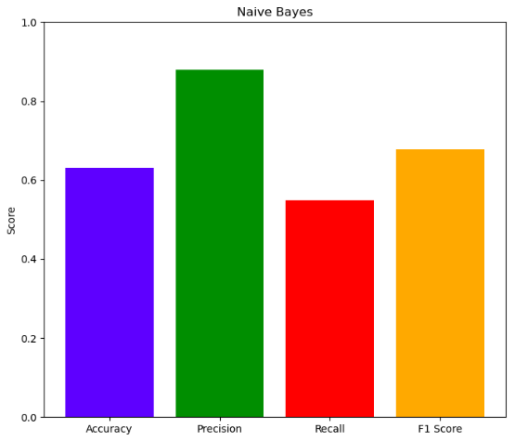

Figure 5: Graph for bayes.

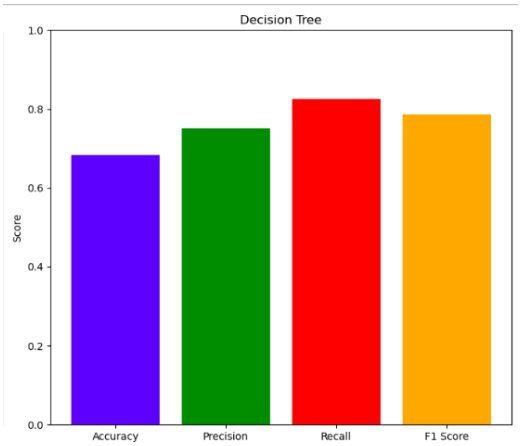

Figure 6: Graph for decision tree.

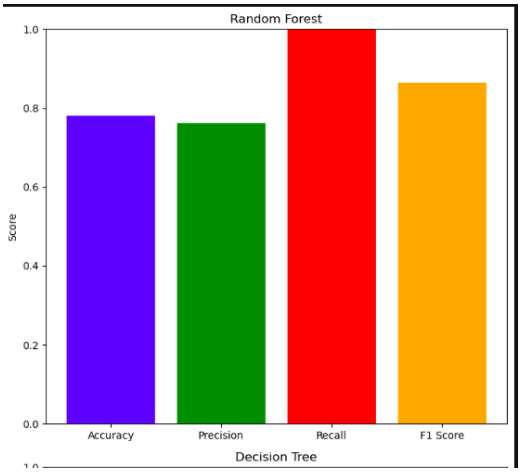

Figure 7: Graph for random forest.

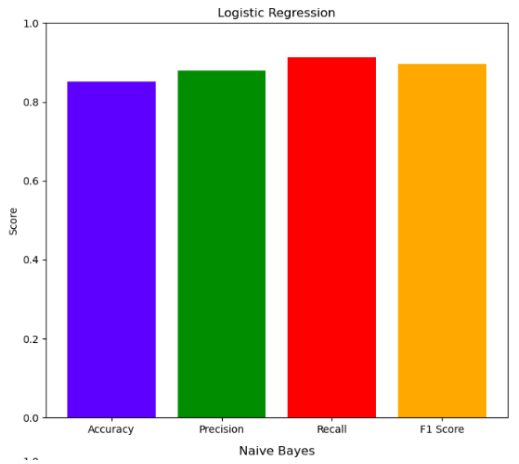

Figure 8: Graph for logistic regression.

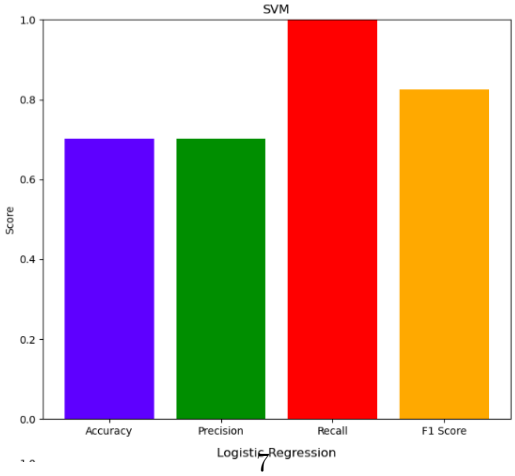

Figure 9: Graph for svm.

Figure 10: Results

5.3 Limitations and Future Directions

Despite the promising results, our study has several limitations that warrant consideration. Firstly, the dataset used for training and evaluation may lack diversity in terms of patient demographics, imaging modalities, and disease severity. Incorporating larger and more diverse datasets from multiple sources could improve the generalizability and robustness of the CNN model.

Additionally, the interpretation of CNN models remains challenging due to their black-box nature, making it difficult to understand the underlying features driving the predictions. Future research efforts could focus on developing interpretable deep learning models or utilizing techniques such as attention mechanisms to identify relevant regions in the images contributing to the classification decisions.

Furthermore, the deployment of the CNN model in real-world clinical settings would require validation on independent datasets and rigorous testing to assess its reliability and performance in diverse patient populations. Collaboration with healthcare institutions and regulatory bodies would be essential to ensure the safe and ethical implementation of automated diagnostic tools for SCD.

In conclusion, our study demonstrates the potential of machine learning, particularly CNNs, in advancing the automated diagnosis of sickle cell disease from medical images. While further research is needed to address the aforementioned limitations and challenges, the development of accurate and efficient diagnostic tools has the potential to improve patient outcomes and enhance healthcare delivery for individuals affected by SCD.

6 Conclusion

In this study, we developed and evaluated a convolutional neural network (CNN) model for the automated diagnosis of sickle cell disease (SCD) from medical images. The CNN model demonstrated high accuracy, precision, recall, and F1 score on the test dataset, indicating its effectiveness in accurately classifying images as either sickle cell positive or negative. Furthermore, the CNN model outperformed traditional machine learning classifiers, including Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and Naive Bayes, highlighting the superiority of deep learning approaches for image-based medical diagnosis.

The success of the CNN model underscores the potential of machine learning technologies in improving the efficiency and accuracy of SCD diagnosis, which is critical for timely intervention and patient management. By automating the diagnostic process, healthcare professionals can expedite treatment decisions, reduce diagnostic errors, and improve patient outcomes, particularly in resource-constrained settings where access to expert clinicians may be limited.

However, our study also revealed several limitations and challenges that warrant further investigation. These include the need for larger and more diverse datasets, the development of interpretable deep learning models, and validation of model performance in real-world clinical settings. Addressing these limitations will be crucial for the successful translation of machine learning-based diagnostic tools into clinical practice.

In conclusion, the findings of this study highlight the potential of machine learning, particularly CNNs, in revolutionizing the diagnosis of sickle cell disease. By leveraging the power of artificial intelligence, we can enhance the accuracy, efficiency, and accessibility of SCD diagnosis, ultimately improving the quality of care for individuals affected by this debilitating condition.

7 References

[1]. H. Frangoul, D. Altshuler, M.D. Cappellini, Y.-S. Chen, J. Domm, B.K. Eustace, J. Foell, J. de la Fuente, S. Grupp, R. Handgretinger, T.W. Ho, A. Kattamis, A. Kernytsky, J. Lekstrom-Himes, A.M. Li, F. Locatelli, M.Y. Mapara, M. de Montalembert, D. Rondelli, A. Sharma, S. Sheth, S. Soni, M.H. Steinberg, D. Wall, A. Yen, S. Corbacioglu, CRISPR-Cas9 gene editing for sickle cell disease and -thalassemia, N. Engl. J. Med. 384 (2021) 252–260, https://doi.org/10.1056/ nejmoa2031054.

[2]. G.J. Kato, F.B. Piel, C.D. Reid, M.H. Gaston, K. Ohene-Frempong, L. Krishnamurti, W.R. Smith, J.A. Panepinto, D.J. Weatherall, F.F. Costa, E.P. Vichinsky, Sickle cell disease, Nat. Rev. Dis. Prim. 4 (2018) 1–22.

[3]. D.S. Charles Bishop, Function and red blood cell, Nature (1965) 435.