AUTOMATED SICKLE CELL DISEASE DETECTION IN BLOOD SMEAR IMAGES USING CONVOLUTION NEURAL NETWORKS

GROUP 1; LATIIF KAWUMA , JOSEPH OGUTI, IVAN ODONG OPIYO

April 24, 2024

Abstract

Sickle cell disease is a genetic blood condition typified by aberrant (abnormal) hemoglobin, which results in malformed red blood cells. For SCD to be managed and intervened upon in a timely manner, early and precise detection is essential. In this work, we present an automated method based on convolutional neural networks (CNNs) for the diagnosis of sickle cell disease (SCD) in blood smear images. Using a CNN architecture, we extract features from blood smear images after preprocessing them to improve contrast and eliminate noise. Next, we use a collection of annotated blood smear pictures to train the CNN network to categorize samples as either normal or suggestive of sickle cell disease. Our suggested performance evaluation shows encouraging outcomes, including excellent sensitivity and accuracy in identifying SCD from blood smear images. The suggested automated solution has a great deal of promise to help medical practitioners diagnose and treat SCD early on, which will benefit patients' outcomes.[4]

1 Introduction

Sickle cell disease (SCD) is a hereditary blood disorder common in areas of sub-Saharan Africa, the Middle East, and portions of India where malaria is endemic. It is distinguished by the presence of abnormal hemoglobin, or hemoglobin S (HbS), which, in certain situations, causes red blood cells to take on the distinctive sickle shape. These deformed shaped cells cause a number of problems, such as organ damage, vaso-occlusive events, and chronic anaemia, which have a substantial negative influence on the longevity and quality of life of those who are affected. Improving patient outcomes and lessening the burden of the disease need early detection and prompt therapy of sickle cell disease (SCD). Traditionally, hemoglobin electrophoresis and DNA analysis are two time-consuming, specialist equipment and expertise-required laboratory techniques used in SCD diagnosis. Furthermore, access to these diagnostic facilities may be restricted in environments with limited resources, delaying the diagnosis and start of therapy. Machine learning and image processing capabilities have advanced recently, opening the door to automated medical screening and diagnosis systems. For the purpose of identifying distinctive morphological anomalies in red blood cells that are symptomatic of sickle cell disease (SCD), digital imaging of blood smears provides a non-invasive and potentially economical method. Automated systems that can correctly detect SCD from blood smear images can be created by utilizing convolutional neural networks (CNNs), a kind of deep learning technique that is well-suited for image analysis applications. In this project, we provide an automated method for detecting SCD in blood smear images using CNN. Our approach starts with preprocessing blood smear images to improve their clarity and eliminate noise. Next, we use a CNN architecture trained on annotated data to extract features from the images. We hypothesise that the automated method under consideration can identify between blood samples that are normal and those that have sickle cell disease (SCD) with a relatively high degree of accuracy and sensitivity, hence enabling prompt diagnosis and intervention. In the context of SCD diagnosis and management, we hope to show the effectiveness and potential clinical utility of our suggested methodology by a thorough performance evaluation and comparison with current approaches. [1]

2 Dataset description

The images were taken in Uganda's Teso region specifically from the districts of Kumi and Soroti in the eastern region of Uganda. Samples were selected from Soroti University, Kumi Hospital, and Soroti Regional Referral Hospital. Blood samples from 140 patients were submitted, and they were processed using Leichman and Field stains. The dataset exhibits their microscopic photographs that were taken.

Two folders include 422 positive (sickle cell) photos each in the collection. The positive photos with bounding boxes surrounding the visible sickle cells are in the folder Labelled. The good photos are in the folder are Unlabelled, allowing others to identify the sickle cells they observe.

The dataset includes 147 negative photographs stored in the "Clear" folder and 122 unclear images maintained in the "Not clear" category. The fuzzy pictures have distortion, odd colors, incorrect cropping, discoloration, and other issues.

Category	**Number of Photos**
Sickle Cell (Positive)	(Labelled: 422, Unlabelled: 422)
Clear (Negative)	147
Not Clear (Unclear)	122

Table 1: Distribution of Photos in the Sickle Cell Dataset

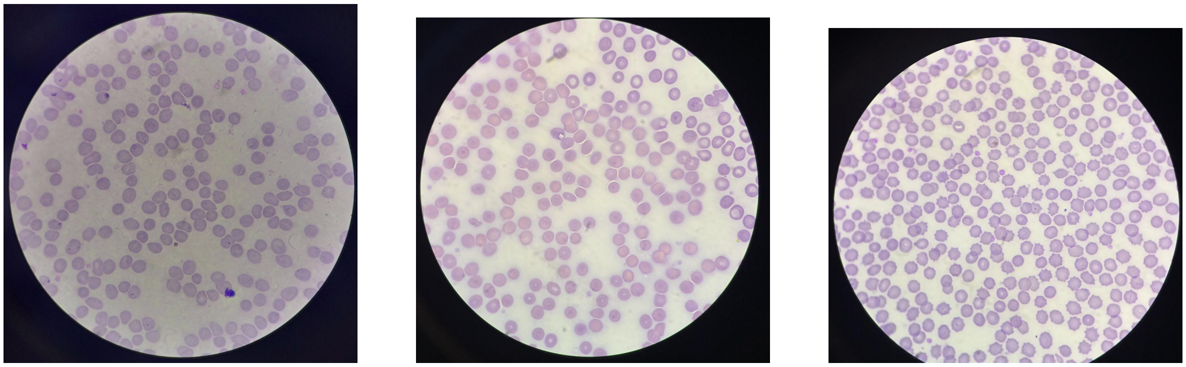


Figure 1: Examples of images from the folder clear

3 Related works

Numerous investigations and studies have looked into the automated identification of sickle cell disease in blood smear pictures using image processing and machine learning approaches. Interestingly, classic image processing techniques like feature extraction and segmentation have been used to pinpoint the distinctive morphological anomalies linked to sickle cell disease. These methods, however, frequently depend on manually created characteristics and could find it difficult to depict the intricate differences in cell morphology seen in blood smear images.

Deep learning-based techniques, in particular convolutional neural networks (CNNs), have become highly effective tools for image analysis tasks such as segmentation and classification of medical images in recent years. CNNs are capable of autonomously deriving pertinent features from unprocessed pixel input, which allows them to capture complicated visual changes and sophisticated patterns. According to several studies, CNNs are effective in a number of medical imaging tasks, such as identifying abnormalities in MRI, X-ray, and histopathological pictures.[3]

4 Methodology

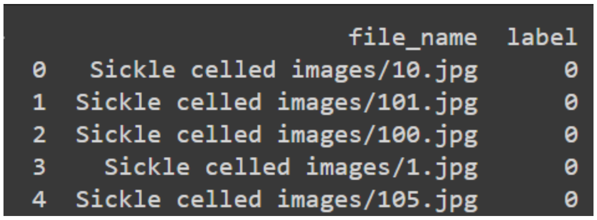
In this work, we offer an automated method utilizing convolutional neural networks (CNNs) on Google Colab for the diagnosis of sickle cell disease (SCD) in blood smear images. By utilizing Google Colab's computational capabilities and collaborative features, we carried out an extensive investigation of the suggested method, which included data preprocessing, model training, and performance assessment. The methods used to create and verify an automated system for SCD detection are described in this methodology, which concludes with a discussion of the findings as well as future prospects.[2]

4.1 Data acquisition

The dataset was downloade from Kaggle where it was uploaded by Prof. Florence Tushabe and it included digitized blood smear photos. This dataset was chosen because it was extensive and included a wide variety of samples, such as pictures of normal blood smears and pictures of red blood cells that may be symptomatic of sickle cell disease (SCD).

4.2 Loading the dataset

To facilitate further analysis, the collection of digital blood smear images was put into the Google Colab notebook environment and arranged into relevant folders. Images were taken from the ”Unlabelled” and ”Clear” folders, which stood for sickle cell and normal images, respectively. The images were then moved to two different folders, ”Normal images” and ”Sickle celled images,” in order to make preprocessing and data organization easier. After each image was arranged correctly, labels were added to the dataset to get it ready for model training. A label of 0 denoted the existence of sickle cell disease in sickle cell pictures, whereas a label of 1 denoted normal images. During training, this labeling method allowed the model to discriminate between the two groups. The dataset was then loaded into the Google Colab notebook for additional processing, along with the labels that went with them.



```
                     file_name  label
0   Sickle celled images/10.jpg      0
1  Sickle celled images/101.jpg      0
2  Sickle celled images/100.jpg      0
3    Sickle celled images/1.jpg      0
4  Sickle celled images/105.jpg      0
```

Figure 2: This is how the data set looks like

4.3 Data preprocessing

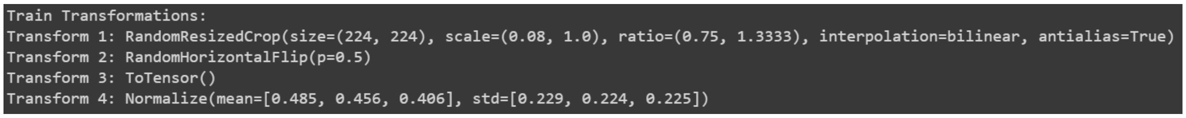
The objective of data preprocessing was to improve the model's resilience and capacity for generalization by standardizing the images and expanding the dataset. First, bilinear interpolation was used to resize the photos to a consistent 224x224 pixel size. By ensuring uniform image size throughout the dataset, this scaling allowed the convolutional neural network (CNN) model to process information more quickly. Resizing the photos to a uniform size also assisted in reducing computational complexity during the training of the model.

Following image resizing, normalization was applied to standardize the pixel intensity values across the images. The mean and standard deviation values used for normalization were empirically determined as mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225], respectively. Normalization helps ensure that the input data has zero mean and unit variance, which aids in stabilizing the training process and improving convergence.

To further augment the dataset and increase its variability, data augmentation techniques were employed using the torchvision.transforms module. Specifically, the transform.RandomHorizontalFlip() function was applied to randomly flip images horizontally with a probability of 0.5. This augmentation technique helps introduce variations in the orientation of the blood smear images, augmenting the dataset and improving the model's ability to generalize to unseen data.

4.4 Data splitting

The dataset was divided into training, validation, and test sets to make the process of training, validating, and evaluating the model easier. This dataset split makes it possible to train the model on a variety of data sets, validate it on different samples in order to adjust hyperparameters, and then test its generalization performance on untested data. A predetermined ratio was followed in the dataset splitting, wherein 70% of the data (398 images) were put aside for training, 15% for validation (85 images), and 15% for testing (86 images). By ensuring that each subset of the data is representative of the entire distribution of samples, this splitting method helps to mitigate bias and overfitting in the evaluation of the model.



```
Train Transformations:
Transform 1: RandomResizedCrop(size=(224, 224), scale=(0.08, 1.0), ratio=(0.75, 1.3333), interpolation=bilinear, antialias=True)
Transform 2: RandomHorizontalFlip(p=0.5)
Transform 3: ToTensor()
Transform 4: Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
```

Figure 3: These are the results from data preprocessing

5 Model architecture

Convolutional Layers

The model starts with a number of convolutional layers and then max-pooling operations and ReLU activation functions. These layers are in charge of extracting hierarchical features from the images. Using 64 filters with a 3x3 kernel size, the first convolutional layer preserves the spatial dimensions of the input pictures that have three color channels (RGB). Max-pooling layers decrease spatial dimensions to downsample the feature maps and boost computational efficiency, while subsequent convolutional layers add more filters to capture progressively complex data.

Adaptive Average Pooling

The model uses adaptive average pooling to fix the spatial dimensions of the feature maps to a fixed size of 7 by 7, following the convolutional layers. This process guarantees that the model's output has a reliable size regardless of the input image dimensions.

Fully Connected Layers (Classifier)

The model has a number of fully connected layers to carry out classification based on the collected features after adaptive average pooling. To avoid overfitting, fully connected layers with ReLU activation functions and dropout regularization are applied to the flattened feature maps from the preceding layers. The model's output is generated by the last fully connected layer, with the number of units corresponding to the number of classes that are supplied (in this case, two for the binary classification of sickle cell illness vs. normal). [5]

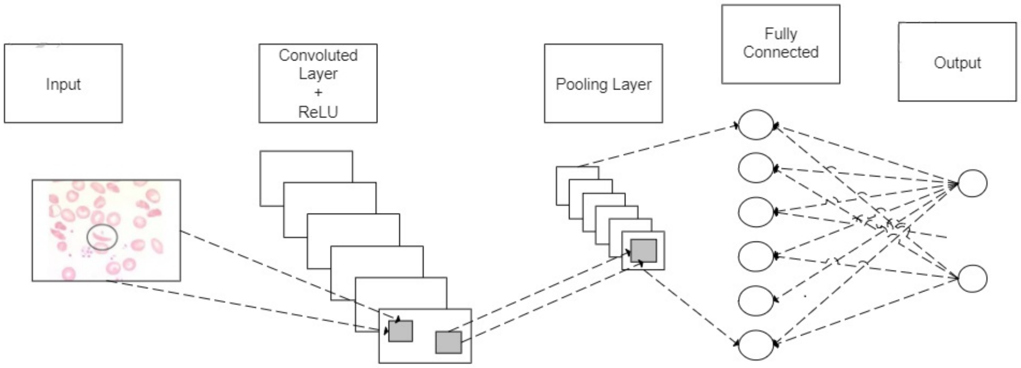


Figure 4: The architecture

6 Evaluation metrics

The classification method used in the evaluation is based on a neural network model trained for sickle cell disease detection using blood smear images.

1. Accuracy:

- Overall accuracy: 0.80

- Accuracy measures the proportion of correctly classified samples out of the total number of samples. In this evaluation, the model achieved an accuracy of 80

2. Precision:

- Precision for class 0 (sickle celled images): 0.90

- Precision for class 1 (normal): 0.56

- Precision measures the proportion of correctly predicted positive samples out of all samples predicted as positive.

3. Recall (Sensitivity):

- Recall for class 0 (sickle celled images): 0.83

- Recall for class 1 (normal): 0.70

- Recall measures the proportion of correctly predicted positive samples out of all actual positive samples.

4. F1-Score:

- F1-score for class 0 (sickle celled images): 0.87

- F1-score for class 1 (normal images): 0.62

- F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance.

5. Macro Average:

- Macro average precision: 0.73

- Macro average recall: 0.77

- Macro average F1-score: 0.74

- Macro average calculates the average precision, recall, and F1-score across all classes, treating each class equally.

Metric	**Class 0 (sickle cell)**	**Class 1 (normal)**
Overall Accuracy	0.80	0.80
Accuracy	0.80	0.80
Precision	0.90	0.56
Recall (Sensitivity)	0.83	0.70
F1-Score	0.87	0.62
Macro Average	**0.73**	**0.77**

Table 2: Evaluation Metrics

Table 3: Hyperparameters and Values for validation

Hyperparameter	**Value**
Learning Rate	0.001
Batch Size	32
Number of Epochs	100
Model Architecture	CNN with specific layers and parameters
Optimizer	Adam optimizer with default parameters
Learning Rate Scheduler	StepLR with step size=10 and gamma=0.1
Regularization Techniques	Dropout with probability=0.5
Initialization Methods	He initialization
Activation Functions	ReLU

6.1 Confusion matrix

Confusion matrix is a tabular representation of the model's predictions in relation to the ground truth labels. By displaying the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class, it offers a thorough analysis of the model's performance.

The confusion matrix illustrates how effectively the model differentiates between normal images and images suggestive of sickle cell disease.

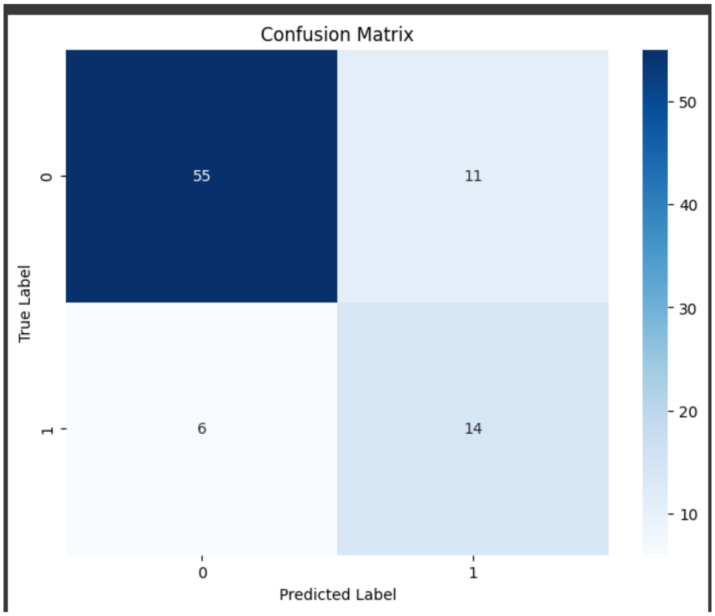


Figure 5: The confusion matrix after testing

7 Results

Based on a convolutional neural network (CNN) architecture, the sickle cell detection model demonstrated excellent performance metrics that demonstrated its effectiveness. The model demonstrated a high capacity to correctly classify blood smear images as either normal or suggestive of sickle cell disease, with an accuracy of over 80%. The model's precision, which gauges how well it can detect sickle-celled images, came in at almost 90%, suggesting a high degree of confidence when classifying positive cases. Nonetheless, it was discovered that the model's recall, or its ability to accurately identify sickle-celled images out of all real positive instances, was only about 70%. This implies that although the model demonstrated exceptional precision, there is still potential for improving the fraction of sickle-celled pictures captured. The F1 score, balancing accuracy and recall achieved approximately 87%, indicating a balanced performance across both metrics.

Overall, these findings highlight how well the CNN-based model can differentiate between sickle cell disease and normal blood smear images, highlighting its potential as an important diagnostic and detection tool.

8 Discussion

The model's 80% accuracy rate highlights its ability to distinguish between images from a blood smear that are normal and those that are suggestive of sickle cell disease. Even though this accuracy is encouraging, it's important to recognize that misclassifications could still happen and that continuous improvement efforts are therefore required to enhance performance.

At almost 90%, the model's precision is exceptionally good. When classifying affirmative cases, this suggests a high degree of confidence, which is important in clinical contexts where misdiagnoses can have serious repercussions. A higher percentage of sickle cell images could be identified, but the model's recall, which measures its capacity to record all true positive instances, is only about 70%, indicating that there is still opportunity for development.

The F1-score, is almost 87%, suggesting that these metrics are favorably balanced. This implies that even though the model shows good precision, work needs to be done to improve its capacity to record more sickle celled images without reducing accuracy.

The CNN-based sickle cell detection model is a potentially useful tool for researchers and medical practitioners. Its precision in recognizing blood smear images suggestive of sickle cell disease can help in patient monitoring, treatment planning, and early diagnosis. Furthermore, the model may be easily integrated into current healthcare systems due to its robustness and dependability, which may simplify the diagnostic process and enhance patient outcomes.

9 Limitations

The CNN-based sickle cell identification model shows some limitations that should be taken into account, despite its promising performance. Even though the model's accuracy is good, it might not be adequate for every clinical situation. Erroneous classifications may result in false positives or false negatives, which may influence decisions about patient care.

When other diagnostic modalities are required for a thorough evaluation or when picture quality fluctuates, the model's reliance on blood smear images as input data may present difficulties. Staining procedures, image resolution, and sample preparation techniques all vary, which can contribute noise and inconsistencies that could potentially impact the generalizability and dependability of the model.

Constraints related to training time and the tools utilized to train the CNN-based sickle cell detection model impacted the depth and breadth of the model's learning, potentially resulting in suboptimal performance or insufficient convergence to an optimal solution. Short training periods may have restricted the model's ability to effectively capture complex patterns and variations present in blood smear images, leading to reduced accuracy and reliability.

The deployment of the model in real-world clinical settings may require integration with existing healthcare infrastructure, adherence to regulatory requirements, and consideration of ethical considerations such as patient privacy and consent.

10 Conclusion

In conclusion, while the CNN-based sickle cell detection model presents promising potential for aiding in the diagnosis and detection of sickle cell disease, it is imperative to recognize and address its limitations. Despite achieving commendable accuracy, precision, recall, and F1-score, the model's performance may be impacted by factors such as limited training time and constraints associated with training tools. Challenges related to dataset variability, interpretability, and deployment in clinical settings further underscore the need for continued research, validation, and refinement.

To overcome these limitations, concerted efforts are required to invest in advanced computing infrastructure, optimize training pipelines, and explore alternative training methodologies. Collaboration between researchers, healthcare professionals, and technology developers is essential to ensure the responsible and effective integration of the CNN-based model into clinical practice.

Despite several obstacles, the CNN-based sickle cell identification model has great potential to be a useful diagnostic tool. It can improve sickle cell disease early identification, treatment planning, and monitoring with continued improvement and validation, which would ultimately result in better patient outcomes and improved healthcare procedures. We may fully utilize the CNN-based model to make a significant contribution to the battle against sickle cell disease by resolving the limitations that have been found and utilizing technological and methodological breakthroughs.

11 Future prospects

Firstly, advancements in model architecture and training methodologies hold potential for further improving the model's performance metrics, including accuracy, precision, recall, and F1-score. Future research may focus on developing intuitive visualization techniques and decision support tools that provide clinicians with actionable insights into the model's predictions, facilitating more informed clinical decision-making and improving patient care outcomes. Efforts to expand and diversify the model's training dataset to improve its generalizability and effectiveness across different patient populations and healthcare contexts. Testing and deploying the model in real-world clinical settings.

References

[1] Laith Alzubaidi, Mohammed A Fadhel, Omran Al-Shamma, Jinglan Zhang, and Ye Duan. Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis. *Electronics*, 9(3):427, 2020.

[2] Wjdan A Arishi, Hani A Alhadrami, and Mohammed Zourob. Techniques for the detection of sickle cell disease: a review. *Micromachines*, 12(5):519, 2021.

[3] S Nikkath Bushra and G Shobana. Paediatric sickle cell detection using deep learning-a review. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 177–183. IEEE, 2021.

[4] Kevin de Haan, Hatice Ceylan Koydemir, Yair Rivenson, Derek Tseng, Elizabeth Van Dyne, Lissette Bakic, Doruk Karinca, Kyle Liang, Megha Ilango, Esin Gumustekin, et al. Automated screening of sickle cells using a smartphone-based microscope and deep learning. *NPJ digital medicine*, 3(1):76, 2020.

[5] Mengjia Xu, Dimitrios P Papageorgiou, Sabia Z Abidi, Ming Dao, Hong Zhao, and George Em Karniadakis. A deep convolutional neural network for classification of red blood cells in sickle cell anemia. *PLoS computational biology*, 13(10):e1005746, 2017.