Abstract—Malaria is a mosquito-borne blood disease caused by the Plasmodium parasites that infect the red blood cells (RBCs). Manual classification and counting of parasitized cells in microscopic thick and thin blood inspection remains the ordinary, but laborious method for disease diagnosis. This project addresses a medical image analysis problem using machine learning techniques, and comparative study with another techniques. It presents a web-based diagnosis tool for malaria parasites species that cause diseases such as P. falciparum, P. vivax, P. ovale, P. Knowlesi, and P. malariae. But P. falciparum can be lethal and infects most of the global population. As the Malaria is an acute febrile illness, especially non-immune individual. Symptoms usually appear 10–15 days after the infective mosquito bite, the first sign of including symptoms, fever, headache, and chills, may be mild and difficult to recognize as The Malaria. If not treated within 24 hours, P. falciparum malaria can progress to severe illness, often leading to death.

In the severe Malaria cases, children develop one or more of the following symptoms, including severe anemia, respiratory distress in relation to metabolic acidosis, and cerebral malaria. In adults, multi-organ failure may also be found. In Malaria endemic areas, people may develop partial immunity, which allows the occurrence of asymptomatic infections.

Diagnosis of direct malaria is done by microscopic blood tests using blood smear or antigen-based rapid diagnostic tests. Methods that use the polymerase chain reaction to detect the parasite's DNA have been developed but is not widely used in areas with general malaria due to expensive and complex procedures. However, WHO recommends all cases of suspected Malaria to be confirmed by parasite-based diagnostic testing, microscopy or rapid diagnostic test, before administering treatment. Parasitological confirmation result can be complete within 30 minutes. Therefore, we propose a web-based parasite detection of malaria by uploading blood smear image into web for malaria diagnosis.

Keywords — Data mining, Machine learning, Malaria, Blood smear, Computer-aided diagnosis.

I. INTRODUCTION

In 2016, the World Health Organization (WHO) reported 212 million instances of the disease across the world [3]. Malaria is a serious and life-threatening disease caused by female anophelines mosquitoes. It causes by the Plasmodium parasitic infection. The parasites grow in the liver, enter the bloodstream and infect red blood cells resulting in serious symptoms. There are several parasites species that cause diseases such as Plasmodium falciparum, P. vivax, P. ovale, P. Knowlesi, and P. malariae. But P. falciparum can be lethal and infects most of the global population. As the Malaria is an acute febrile illness, especially non-immune individual. Symptoms usually appear 10–15 days after the infective mosquito bite, the first sign of including symptoms, fever, headache, and chills, may be mild and difficult to recognize as The Malaria. If not treated within 24 hours, P. falciparum malaria can progress to severe illness, often leading to death.

In the severe Malaria cases, children develop one or more of the following symptoms, including severe anemia, respiratory distress in relation to metabolic acidosis, and cerebral malaria. In adults, multi-organ failure may also be found. In Malaria endemic areas, people may develop partial immunity, which allows the occurrence of asymptomatic infections.

Diagnosis of direct malaria is done by microscopic blood tests using blood smear or antigen-based rapid diagnostic tests. Methods that use the polymerase chain reaction to detect the parasite's DNA have been developed but is not widely used in areas with general malaria due to expensive and complex procedures. However, WHO recommends all cases of suspected Malaria to be confirmed by parasite-based diagnostic testing, microscopy or rapid diagnostic test, before administering treatment. Parasitological confirmation result can be complete within 30 minutes. Therefore, we propose a web-based parasite detection of malaria by uploading blood smear image into web for malaria diagnosis.
two-dimensional data such as images and videos. In a ten-fold cross-validation based on 27,558 single cell images, the average accuracy of their new 16-layer CNN model is 97.37%. However, that they use a convolutional neural network (CNN) model when they have many layers, each training step is going to take much longer.[4]

III. MATERIAL AND METHODS

In this project, we propose a machine learning model to classify malaria parasite cells of humans from Malaria Datasets.[1]

The objective of this comparative research work is to investigate various machine learning techniques to identify more suitable ML techniques with favorable classification results and develop a web-based diagnosis tool for malaria parasites detection.

This classification of dataset already exists, but the National Library of Medicine and most people use deep learning to make the classification.

So, we are interested in developing a machine learning model that will be used as an algorithm for building a tool that will reduce predictive time that is as effective as deep learning and can be used as a client-side only.

This research work is designed based on CRISP-DM methodology. CRISP-DM is an open standard process model that describes common approaches used by data mining experts. It is the most widely used analytics model.[5]

![CRISP-DM](image)

**Figure 1 CRISP-DM.**

A. Business Understanding

Microscopic thick and thin blood smear examinations are the most reliable and commonly used method for Malaria diagnosis. Thick blood smears assist in detecting the presence of parasites while thin blood smears assist in identifying the species of the parasite causing the infection [3].

The diagnostic accuracy heavily depends on human expertise and can be adversely impacted by the inter-observer variability and the liability imposed by large-scale diagnoses in disease-endemic or resource-constrained regions [1].

In our works, we presented a data classification model for malaria diagnosis by Consider whether red blood cells are infected with P. falciparum parasites using a Machine Learning Techniques.

B. Data Understanding

To reduce the burden for microscopists in resource-constrained regions and improve diagnostic accuracy, researchers at the Lister Hill National Center for Biomedical Communications (LHNBC), part of National Library of Medicine (NLM), have developed a mobile application that runs on Android smartphone that can be attached to a conventional light microscope [1].

Giemsa-stained thin blood smear slides from 150 P. falciparum-infected and 50 healthy patients were collected and photographed at Chittagong Medical College Hospital, Bangladesh.

The smartphone’s built-in camera acquired images of slides for each microscopic field of view. The images were manually annotated by an expert slide reader at the Mahidol-Oxford Tropical Medicine Research Unit in Bangkok, Thailand. The de-identified images and annotations are archived at NLM.

The dataset contains a total of 27,558 cells images with equal instances of parasitized and uninfected cells.

C. Data Preparation

We have the main process for data preparation include the following processes.

1) **Image scaling & Uniform aspect ratio**

Images in the dataset had various sizes, therefore images had to be resized before being used as input to the model. All the images were resized to 100x100 pixels (10,000 pixels) by Open Source Computer Vision Library (OpenCV). OpenCV is an open source computer vision and machine learning software library.

2) **Dimensionality reduction**

And convert each pixel from 3 channels (R, G and B) to 1 channel decimal values using the following formula and, convert images matrix to CSV file. We used Pandas for python to be done this. Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for Python programming language.

\[
RGB = 256^2R + 256G + B
\]

(1)

3) **Data Normalisation**

All the image input features (Pixel, in this case) are based on 10,000 features for each image. However, each of images has different scales. The data normalization is used to scale all of features between 0 and 1 using following formula, we used Weka software. This makes convergence faster while training.

\[
\text{Feature} = \frac{\text{Value} - \text{Min Value}}{\text{Max Value} - \text{Min Value}}
\]
\[
z = \frac{x - \min(x)}{\max(x) - \min(x)}
\]

4) **Feature selection using Information Gain**

We use information gain to identify the optimal split. The information gain “information content” in the data partition. After, we received the data normalization, it will enter information gain process by using Weka software. The results of information gain can tell which pixels are affecting more or less. The expected information needed to classify a tuple in D is given by

\[
\text{Info}(D) = - \sum_{i=1}^{m} P_i \log_2(P_i).
\]

(3)

How much more information would we still need (after the partitioning) to arrive at an exact classification? This amount is measured by

\[
\text{Info}_A(D) = \sum_{j=1}^{p} \frac{|D_j|}{|D|} \times \text{Info}(D_j).
\]

(4)

Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

\[
\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D).
\]

(5)

In other words, Gain(A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A.[6]

5) **Convert image to HSV & Histogram**

We changed the color space from BGR to HSV in order to create a new image with color space that were initially similar to more distinct colors. In order to clearly see where there is a part of the image that is to be blamed. Then convert the HSV image to histogram to find the frequency of channel 1 (Green). As we have observed that the image obtained from BGR to HSV, the area of the defective image will look green, allowing it to classify easier. This is the additional processes that we have used to improve the efficiency of the model. We used 0 and 1 for classes label, 0 for Parasitized and 1 for Uninfected.

D. **Modeling**

In a CRISP-DM methodology, after data preparation, different classification algorithms are employed to classify the data. In this study, we have considered two widely used algorithms for this study we decided to use Decision Trees (DT) algorithm because it is a model that can be used to make classification model and is a model that can be easily understood. In technical, the Sci-kit-learn library package was used in Python programming language for the implementation.

1) **Decision tree with full features**

Firstly, we train dataset with decision tree using all features of dataset (all pixels) then we and then we get an accuracy of around 0.612, which we think is quite low for classification. So, we want to find a technique that will increase the accuracy of this algorithm model.
2) Decision tree according to rank obtained from information gain with step by 10 features to full features

In this process, we run the model by running at 10 features in increments of 10 to running at 20 features, 30 features and so on until we run with full features (10,000 features) in order according to the rank received from the information gain. And take the accuracy, precision, recall and f1-score values to store in CSV and visualize them in graph. There will be differences in every run and every period. Even if not running at full features can give accuracy and other values than running at full features. Therefore, this feature selection will give better results.

3) Decision tree with converting to histogram technique at full features

After we run DT and got the result but we were not satisfied with the result because the highest accuracy was only 0.678, we added the process to convert BGR color space to HSV color space and convert the HSV image to histogram and we use this technique with full feature of image.

![Data Preparation and Modeling design](image)

*Figure 4 Data preparation and Modeling design.*

![Accuracy graph](image)

*Figure 5 Accuracy obtained from running DT according to rank obtained from information gain step by 10.*
Figure 6 Precision obtained from running DT according to rank obtained from information gain step by 10.

Figure 7 Recall obtained from running DT according to rank obtained from information gain step by 10.
4) **Decision tree according to rank obtained from information gain with converting to histogram technique and features selection**

At this process, we combine the techniques from items 2 and 3 together, in which we will select feature 8450 feature based on the result of information gain obtained from Weka software. The reason we choose the number of such feature is because we have very accurate. Finally, at 8540 features from item 2, then combined with item 3, convert RGB color space to HSV color space and convert the result to histogram.

![Partition of Decision Tree Structure](image)

**E. Evaluation**

This classification model is evaluate based on unbiased methods. Generally, the original dataset is separated into a training set and test set. The training set is evaluate based on the k-fold cross-validation techniques. It is resampling process that is widely used to evaluate ML models. A k must be specified. The k refers to number of groups that must be split. The original dataset must be randomly separated into k group. For example, if k is specified as 10, ten separated group are created from the original dataset. Cross-validation is mostly used to evaluate overall performances of ML model. Finally, a selected model is used to classify the test set. The separated unseen test set reduces the bias to the model.

We separated dataset to unseen set to test in real-world situation for 20%. Another 80% is training set was divided into 10 groups for training set 9 groups and another 1 group for validation set. We train model by Cross-validation with shuffle data in each group for more various of learning.

![F1-score graph](image)
For the performance evaluation, measurements have been selected each of the techniques according to the sleep stages. The measurements include Accuracy, Precision, Recall and F1-score. The accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. Finally, the F1-score is the weighted average of Precision and Recall, which is one of the standard comparative measurements. These selected measurements can be calculated from incremental counts of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) in confusion matrices.

**F. Deployment**

We deployed the trained model into a web browser to enable running the model at reduced computational cost and alleviate issues due to the complex backend. We used JavaScript to develop client-side-only app that using web browser execute model within code, and named the application as Diagnosis Malaria Detection to which the user submits an image of the parasitized/uninfected cells and the model embedded into the browser gives the predictions.

**IV. RESULTS AND DISCUSSION**

The two classes considered were Parasitized and Uninfected. In the dataset 13,779 images were Parasitized, and 13,779 images were Uninfected. Classification were compared in 5 techniques; consisted of Convolutional Neural Network (CNN) of related literature, Decision tree (DT), DT with feature selection (8540 features), DT with histogram and finally Decision tree with histogram and features selection techniques (8540 features). All the techniques were tested in four classifiers for 10 folds.
**A. Performance metrics**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Training Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN *</td>
<td>0.940</td>
<td>0.951</td>
<td>0.931</td>
<td>0.941</td>
<td>15,572.448</td>
</tr>
<tr>
<td>DT (full features)</td>
<td>0.612</td>
<td>0.613</td>
<td>0.612</td>
<td>0.612</td>
<td>1,643.112</td>
</tr>
<tr>
<td>DT (selected 8,540 features)</td>
<td>0.678</td>
<td>0.679</td>
<td>0.678</td>
<td>0.678</td>
<td>974.314</td>
</tr>
<tr>
<td>DT + Histogram (full features)</td>
<td>0.948</td>
<td>0.935</td>
<td>0.964</td>
<td>0.949</td>
<td>12.348</td>
</tr>
<tr>
<td>DT + Histogram (selected 8,540 features)</td>
<td>0.949</td>
<td>0.943</td>
<td>0.955</td>
<td>0.949</td>
<td>9.877</td>
</tr>
</tbody>
</table>

Table 1: Performance metrics of 5 techniques comparison.

Based on the results provided by Table 1, among the five classification techniques, DT with histogram and feature selection (8,540 feature) provides the highest accuracy. However, highest precision were computer-aided diagnostic tools based on data-driven deep learning algorithms like convolutional neural network (CNN). [2]

**B. Discussion and Problem**

A histogram is a unique bar graph that shows the relationship between data in categories called data layers and data frequencies. To see the distribution of information. The characteristics of categorized data are arranged in ascending order, with the number of categories of information classified as appropriate. So, we then convert the HSV image and transform it into a histogram to determine the color channel frequency. And the results obtained It turns out that it was better than we expected.

For the future work, an extensive study is needed to investigate other ensemble techniques, and other CNN classification architectures to be combined with our techniques in the malaria parasite classification task. And we expect that we will get good results as an alternative way to classify malaria parasites.

**V. Conclusions**

Related literature from the discussion of the potential of Convolutional Neural Network (CNN) techniques for identifying malaria data sets which will have more detail and high accuracy, but because their technique is CNN, it takes some time to process.

By the way, we used Decision tree for the classification of Malaria. When the dataset was 10-fold cross-validated and tested, the maximum accuracy was found to be obtained by Decision tree with Histogram and feature selection techniques (8,540 feature). The accuracy obtained was 94.9% and takes less time. Thus in our approach, we have used this technique to obtain the best possible results for classifying Malaria parasites. Therefore, this technique is an alternative way to classify malaria parasites.

**ACKNOWLEDGMENT**

We would like to thank Asst. Prof. Dr. Thakerng Wongsirichot, Information and Communication Technology for support us. We have to express out appreciation to the Mr. Hassan H., Mr. Teepakorn J., Mr. Tossapon K., Mr. Jirapas K. and Mr. Chayanon C. for sharing their computer with us during the course of this research. And the last thank you for all encouraging to project successes.
Figure 12 Snapshot of the web application interface with classified as Parasitic cell.

Figure 13 Snapshot of the web application interface with classified as Normal cell.
REFERENCES


