

Malaria Detection using Image Processing and Machine Learning

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering.

By

Ch HariKalyan Gowtham (38110192)
Venkata Sai Sandesh(38110637)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119

MARCH - 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of
**Ch HariKalyan Gowtham (38110192) Venkata Sai Sandesh
(38110637)** who carried out the project entitled under my supervision
from **June 2021** to **March 2022**.

Internal Guide

DR B ANKAYARKANNI, M.E., Ph.D.,

Head of the Department

DR. VIGNESHWARI S, M.E., Ph. D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **Ch HariKalyan Gowtham (38110192) Venkata Sai Sandesh (38110637)** hereby declare that the Project Report entitled “**Malaria Detection using Image Processing and Machine Learning**” done by me under the guidance of Dr B Anayarkanni(Internal) is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering / Technology degree in **Computer Science and Engineering**.

DATE:

PLACE:

SIGNATURE OF CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr.T.Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr.S.Vigneshwari M.E., Ph.D.**, and **Dr.L.Lakshmanan M.E., Ph.D., Heads** of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.B.Ankayarkanni M.E., Ph.D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

Abstract:

Malaria is mosquito-borne blood disease caused by parasites of the genus Plasmodium. Conventional diagnostic tool for malaria is the examination of stained blood cell of patient in microscope. The blood to be tested is placed in a slide and is observed under a microscope to count the number of infected RBC. An expert technician is involved in the examination of the slide with intense visual and mental concentration. This is tiresome and time consuming process.

In this paper, we construct a new image processing system for detection and quantification of plasmodium parasites in blood smear slide, later we develop Machine Learning algorithm to learn, detect and determine the types of infected cells according to its features

Chapter 1

INTRODUCTION:

The study proposes an image processing model for detection of malaria infected cells. We use image processing techniques to detect parasite-infected red blood cells in thin smears on standard microscope slides. The most widely used present day method is examining thin blood smears under a microscope, and visually searching for infected cells. A clinician manually counts the number of parasitic red blood cells - sometimes up to 5,000 cells (according to WHO protocol). Malaria could be forestalled, controlled, and relieved all the more adequately if an increasingly precise and effective symptomatic techniques were accessible. We have utilized image processing procedures to identify the nearness of malaria contaminated cells. And to classify the stage of malaria i.e falciparum or non-falciparum we use machine learning.

Hundreds of millions of blood films are examined every year for malaria, which involves manual counting of parasites and infected red blood cells by a trained microscopist. Accurate parasite counts are essential not only for malaria diagnosis. They are also important for testing for drug resistance, measuring drug-effectiveness, and classifying disease severity. However, microscopic diagnostics is not standardized and depends heavily on the experience and skill of the microscopist. It is common for microscopists in low-resource settings to work in isolation, with no rigorous system in place that can ensure the maintenance of their skills and thus diagnostic quality. This leads to incorrect diagnostic decisions in the field. For false-negative cases, this leads to unnecessary use of antibiotics, a second consultation, lost days of work, and in some cases progression into severe malaria. For false positive cases, a misdiagnosis entails unnecessary use of anti-malaria drugs and suffering from their potential side effects, such as nausea, abdominal pain, diarrhea, and sometimes severe complications.

Chapter 2

LITERATURE REVIEW

2.1 Malaria Detection using Image Processing and Machine Learning

Abstract

Malaria is a deadly, infectious and lifethreatening mosquito-borne blood disease caused by Plasmodium parasites. The conventional and most standard way of diagnosing malaria is by visually examining blood smears via microscope for parasite-infected red blood cells under the microscope by qualified technicians. This method is inefficient and time consuming and the diagnosis depends on the experience and the knowledge of the person doing the examination. Automatic image recognition technologies based on image processing have been applied to malaria blood smears for diagnosis before. However, the practical performance has not been up to the mark so far. This gives us all the motivation to make malaria detection and diagnosis fast, easy and efficient. Our main aim is to build a model that can detect cells from images of multiple cells in thin blood smear on standard microscope slides and classify them as either infected or uninfected with early and effective testing using image processing. And also perform classification on the infected cell image using machine learning.

INTRODUCTION

Malaria is a deadly, infectious disease caused by the Plasmodium parasite which is transmitted by the bites of female Anopheles mosquitoes. According to the World Malaria Report 2019 published by WHO, there were an estimated 405,000 malaria related deaths in the preceding year. The disease is curable but early detection holds the key. Existing methods used to detect Malaria include microscopic detection of infected cells in a laboratory. The method is both expensive and tedious. An estimated 93 percent of all Malaria cases in 2018 were reported in the WHO African region. The region also has one of the lowest per capita incomes across the world. A faster, low cost, and reliable alternative to microscopic detection of Malaria is proposed in this model.

Problem Statement

We propose an image processing model for detection of malaria infected cells. We use image processing techniques to detect parasite-infected red blood cells in thin smears on standard microscope slides. The most widely used present day method is analyzing thin blood smears under a microscope, and visually searching for contaminated cells. A clinician manually counts the number of parasitic red blood cells - sometimes up to 5,000 cells (according to WHO protocol). Malaria could be forestalled, controlled, and relieved all the more adequately if an increasingly precise and effective symptomatic techniques were accessible. We have utilized image processing procedures to identify the nearness of malaria contaminated cells. And to classify the stage of malaria whether it is falciparum which is the most deadliest stage in malaria or non-falciparum, for this we use machine learning technologies.

Scope of the Project

The Malaria Detection from thin film blood smear images demands segmentation of single blood cells from the microscopic blood slide images which can be taken from a pathologist and the dataset would contain cell images that are not segmented. Hence, segmentation in the proposed method is done using a variety of image processing techniques. Edge detection techniques and segmentation techniques used in this system overcomes the issue of overlapping of cells by eliminating the noise and finding the discontinuities of the cells. It differentiates each cell and detects the infection in the cell using morphological segmentation. Also, all the images are raw and have different intensities, and since there is no uniformity in all the images, detection of cells and infection is very difficult. To overcome this problem, the proposed method uses histogram matching where all the images are standard and has the same intensity which in turn increases the accuracy level.

METHODOLOGY

At present, the recognition of Malaria parasite in single cell slide is totally manual. This procedure could be rearranged by capturing an image of the blood smear and afterwards utilizing the proposed model to arrange whether the cells are contaminated or not. The proposed model uses the utilization of image processing systems to improve existing techniques and abbreviate the time taken for recognition of malaria parasite in blood tests. The dataset is manually collected from the CDC's Division of parasitic infection and Malaria. Our concentration here is to make a mechanized capacity to distinguish the nearness of Malaria parasite in slight blood spread and measure the segment of RBC in the example that are tainted, primary assignment is to fragment the contaminations, for which segmentation of the cells is the earlier undertaking. Segmentation techniques involve methodologies based on Edge Detection, Watershed Segmentation and Morphological segmentation. Once the cells are segmented, the infections are segmented. This is done via using a threshold intensity pixel value for the infection, i.e. if the pixel value is in the range of the threshold the infections are identified.

Segmentation of cells

When we "segment" an image, we distinguish the regions of interest (ROIs) from the non-ROI portion, generally creating a binary mask of what we want to qualify, quantify, track, etc. Segmentation is a critical part of many image processing problems, and is worth considering in some depth. Here also we will segment our cell images. One of the major issues in segmenting human blood cells is the contiguity of the cells, one may find a lot of overlapping of cells. Segmenting touching object is one of the most difficult task in image processing. We use different approaches like Edge Detection, Watershed Segmentation, Morphological Segmentation.

- 1) Edge Detection(Gradient Based Techniques): The gradient based edge detections look for the first derivative of an image where the maxima and minima are occur. These techniques used sobel, prewitt and robert's cross operator for finding the edges.
- 2) Edge Detection(Gaussian Based Techniques): The main purpose of this technique is to detect the zero crossings

in the second order derivative of an image to find edges. The Gaussian based techniques are Laplacian of Gaussian (LOG) and Canny edge detection.

The LOG operator works best in all of the edge detection techniques since it can eliminate the noise levels and discontinuities in the cells better as compared to the other edge detection techniques which would otherwise not prove to be good for further detection and it would be hard to segment the cells.

3) Watershed Segmentation: Watershed segmentation algorithm is applied on the images in the dataset, it segments or divides the different cells present in the image. These divisions help in simplifying the detection and classification of malaria parasite in the blood samples.

4) Morphological Segmentation: The term morphology refers to the shape and size of the objects within the image. In our study we have used the shapes of the cells, as most of the cells were circular in shape and of a particular radii range. Thus, a circle detection algorithm was designed which could detect the cells based on the shape of the cells.

Segmentation of infection

After successful detection of the cells the next task is to spot the infections inside the cells. So, for this we would be using the pixel intensity values, and observe the range of intensities wherever the infections. But the problem with it is that the illumination and color scale of one image may differ to other. For this we would use histogram matching to have a uniform color scale.

Classification of infected cells

We use machine learning to classify the cells if infected. Support Vector Machine(SVM) is a supervised machine learning algorithm that can be used for both classification and regression tasks. An SVM classifies data by finding the best hyperplane that separates data points of one class from those of the other class. The best hyperplane for an SVM implies the one with the largest margin between the two classes. Margin means the maximal width of the slab corresponding to the hyperplane that has no interior data points.

Data Source

We used archived blood smear images acquired from the CDC's Division of parasitic infection and Malaria as input. Our data set contains images of different malaria parasites like Plasmodium Falciparum, Plasmodium Knowlesi, Plasmodium Malariae, Plasmodium Ovale, Plasmodium Vivax.

Steps used in Image Processing

1. We apply an auto-generated segmenter that performs five steps:
 - (1) Convert to grayscale
 - (2) Initialize segmentation with Otsu's threshold.
 - (3) Filter components by area.
 - (4) Form masked from input image and segmented image.

2. Detecting edges-We have used various edge detection techniques like Canny, Laplacian of Gaussian ,Prewitt, Roberts, Sobel and Zerocross operator out of which LOG gives us the best result.
3. Combining the edges(logically) with the segmented regions.
4. Improve edge mask by performing morphological operations like imclose, skeleton, etc.
5. Cleaning of the mask and refine the mask.
6. Use different segmentation techniques like watershed, thresholding, k-means clustering based segmentation. Of which we select watershed segmentation since the other techniques have tiny pores and do not give good results for further segmentation.
7. To improve the result and flatten the pools from the results of watershed we use imhmin function.
8. We use another segmentation technique i.e. Morphological segmentation to overcome the shortcomings in watershed segmentation. This technique detects the circles by setting a specific range of radius.
9. Histogram matching is done since all the images have different intensities and for classification we require all the images to be standard with the same intensity.
10. Finally, if the intensity values are within the threshold we display the percentage of cells infected.

Classification Stage

We used Classification Learner to automatically train a selection of different classification models on our data. We use automated training to quickly try a selection of model types, then explore promising models interactively. Here, the two classes are falciparum and non-falciparum. If the cell is infected in the image processing stage, then we undergo a test for classifying whether the malaria parasite is falciparum or some other malaria parasite(non-falciparum). Machine learning algorithms like Cubic SVM, Linear SVM and cosine KNN are used and out of which the best algorithm with the highest accuracy score is selected for classification purpose.

Outputs

All the images are read in MATLAB and an objective/target image is chosen on which all the tasks are performed and afterward on all the images. After performing all the steps of image processing, the images of blood cells are displayed whether infected or not infected, and if infected the number of cells infected and their percentage is shown and then we train the model for classification of the infected cell.

CONCLUSIONS

Considering our above methodology, we come to the conclusion that the newly designed system of image processing technologies is suitable for parasite detection. The morphological segmentation technique and improved watershed segmentation technique proves to be the best in detection of the cells and segmenting them from the nonregion of interest portion. To characterize the sort of plasmodium parasite for which we utilized machine learning technologies and cubic SVM proves to be the best by having the highest accuracy score. The system has been tested with around 110 thin film blood smear images and the results are satisfactory. The inconsistencies among the images was a challenge but we have tried to make our system robust.

2.2 Image analysis and machine learning for detecting malaria

INTRODUCTION

Malaria is caused by protozoan parasites of the genus *Plasmodium* that are transmitted through the bites of infected female *Anopheles* mosquitoes and that infect the red blood cells. Most deaths occur among children in Africa, where a child dies almost every minute from malaria, and where malaria is a leading cause of childhood neuro-disability. According to the World Malaria Report 2016, an estimated 3.2 billion people in 95 countries and territories are at risk of being infected with malaria and developing disease, and 1.2 billion are at high risk (>1 in 1000 chance of getting malaria in a year). There were about 214 million cases of malaria globally in 2016 and about 438,000 malaria deaths. The burden was heaviest in the African region, where an estimated 92% of all malaria deaths occurred, and in children aged under 5 years, who accounted for more than two thirds of all deaths (see also the malaria death rates from an earlier WHO report. Typical symptoms of malaria include fever, fatigue, headaches, and, in severe cases, seizures and coma, leading to death. Hundreds of millions of blood films are examined every year for malaria, which involves manual counting of parasites and infected red blood cells by a trained microscopist. Accurate parasite counts are essential not only for malaria diagnosis. They are also important for testing for drugresistance, measuring drug-effectiveness, and classifying disease severity. However, microscopic diagnostics is not standardized and depends heavily on the experience and skill of the microscopist. It is common for microscopists in low-resource settings to work in isolation, with no rigorous system in place that can ensure the maintenance of their skills and thus diagnostic quality. This leads to incorrect diagnostic decisions in the field. For false-negative cases, this leads to unnecessary use of antibiotics, a second consultation, lost days of work, and in some cases progression into severe malaria. For false positive cases, a misdiagnosis entails unnecessary use of anti-malaria drugs and suffering from their potential side effects, such as nausea, abdominal pain, diarrhea, and sometimes severe complications. This sober analysis of malaria diagnosis has prompted efforts to perform malaria diagnosis automatically. Automatic parasite counting has several advantages compared with manual counting:

- (1) it provides a more reliable and standardized interpretation of blood films,
- (2) it allows more patients to be served by reducing the workload of the malaria field workers, and
- (3) (3) it can reduce diagnostic costs. Several key processing steps are typically required to quantify parasitemia automatically. First, digital blood slide images need to be acquired, which often requires preprocessing to normalize for lighting or staining variations. In a second step, blood cells or parasites need to be detected. For blood cells, this typically implies cell segmentation to identify individual cells in cell clumps to obtain accurate cell counts. In a third step, after cell detection and segmentation, features are computed to describe the typical visual appearance of infected and uninfected blood cells. In a final classification step, a classifier, who has been trained on an independent and typically manually annotated training set, then discriminates between infected and uninfected cells. Once the number of infected and uninfected cells is known, computation of parasitemia is a straightforward athematical equation, which includes clinical parameters such as hematocrit value, for example. The prospects of automating malaria diagnosis with its obvious advantages has attracted many researchers, especially in the last decade. The publications reflect all

the major developments we have seen in the areas of automatic pattern recognition and machine learning in the last years. Our article will give an overview of the articles that have been published, using the processing steps mentioned above as a framework and guide. This is not the first survey article on the subject. In fact, several survey articles have already been published before, which bear testimony to both the importance of automated malaria diagnosis and the research dynamics and rapid system development. We refer readers in particular to the following surveys for additional information about the background of automatic malaria diagnosis and the image processing and machine learning methods used for automated microscopy diagnosis of malaria. In addition, more specific surveys have been published on cell features for malaria parasite detection, on malaria diagnosis, on malaria diagnostic tools, and on alternatives to conventional microscopy. The purpose of our article is not to replace these surveys, but rather to complement them and to provide the latest update of the state of the art in image analysis and machine learning for malaria diagnosis as it presents itself at the end of the year 2017. With about 170 literature citations, we have collected more references compared with the other surveys. We had the goal to include also maybe lesser known publications to provide a historical documentation of the work done. In addition, we included a section on deep learning, which is the latest development in malaria diagnosis and which arguably has the potential to render many of the old approaches obsolete, similar to the development in other imaging application areas. There have also been many developments in hardware for automatic malaria diagnosis, which are however out of the scope of this article and deserve a separate article. Nevertheless, we devote a section to rapid diagnostic tests (RDTs) for malaria diagnosis because they are also widely used in the field. The bulk of our articles have been collected from the *Journal of Microscopy*, *Malaria Journal*, and *PLOS ONE*, including a few articles from *Nature* and others. We have also collected publications from Institute of Electrical and Electronics Engineers (IEEE) conferences and other proceedings published by Springer and Elsevier. Furthermore, we have organized the articles into sections for preprocessing, cell detection and segmentation, feature computation, and classification. We have also added a separate section about deep learning and an extensive section about mobile smartphone applications for malaria diagnosis. A discussion of the latest developments and our conclusion mark the end of this article.

MALARIA

There are 5 *Plasmodium* species that cause malaria in human: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale*, and *Plasmodium knowlesi*. The 2 most common species are *P. falciparum* and *P. vivax*. *P. falciparum* is the most severe form and is responsible for most malaria-related deaths globally.¹ *P. falciparum* is the most prevalent malaria parasite in sub-Saharan Africa, accounting for 99% of estimated malaria cases in 2016. Outside of Africa, *P. vivax* is the predominant parasite in the WHO Region of the Americas, representing 64% of malaria cases, and is above 30% in the WHO Southeast Asia and 40% in the Eastern Mediterranean regions. Each of these parasite species goes through stages during their development cycle (48 hours), which gives the parasites a different visual appearance that can be observed under the microscope. In chronologic order, these stages are the ring stage, trophozoite stage, schizont stage, and gametocyte stage. typical examples of all stages for each species. In nonsevere malaria, mostly the young stages (<24

hours old) of *P. falciparum* are present in the peripheral blood, whereas for severe malaria all stages can be present in the peripheral blood. For *P. falciparum*, the trophozoite-infected red blood cells disappear from the peripheral blood circulation by attachment to the walls of capillaries inside vital organs, which is a process called sequestration. If the capillaries are blocked for newly infected cells by already attached cells, more mature parasite stages (trophozoites and schizonts) will be visible in the peripheral blood, which indicates a severe infection and a bad prognosis. For *P. falciparum*, ring stages have a visible cytoplasm and 1 or 2 small chromatin dots. The infected blood cells are not enlarged but can feature multiple infections. *P. falciparum* trophozoites are rarely seen in peripheral blood smears. The cytoplasm of mature trophozoites tends to be more dense than younger rings, trophozoites can appear round in shape with brown malarial pigment inside, (Centers for Disease Control and Prevention (CDC)). *P. falciparum* schizonts are also seldomly seen in peripheral blood. They are displaying more than 2 and up to 32 nuclei (merozoites) with dark brown pigment clumped in the middle. Gametocytes of *P. falciparum* have a crescent or sausage shape, and can be seen in the blood smear 1 week after a parasite infection. The chromatin is visible as a single mass or is diffuse. For more information about *P. falciparum* morphology, see for example Similar observations can be made for the stages of the other parasite species. For example, for *P. vivax*, host cells are often enlarged and have irregular shape. Trophozoites are amoeboid in shape with malarial pigment seen, and for severe infections multiple infections of single blood cells are not uncommon. For *P. malariae*, host cells are not enlarged. Trophozoites have a strong tendency to form a band with malarial pigment scattered along across the diameter of infected red blood cells. Multiple infections are extremely rare for *P. malariae*. On the other hand, for *P. ovale*, host cells are slightly enlarged and have an oval shape with tufted ends, often fimbriated. Parasites are slightly enlarged and trophozoites are amoeboid in shape with malarial pigment. Multiple infections of a single cell are more common than for *P. vivax*. For *P. knowlesi*, infected red blood cells do not appear enlarged. The parasite erythrocytic cycle is only 24 hours, which is shorter than *P. falciparum*'s cycle (48 hours) and much shorter than *P. malariae*'s cycle (72 hours), which will lead to the same stage seen in peripheral blood every day at a given time. The morphology of *P. knowlesi* parasites is similar to *P. malariae*. Trophozoites can feature malarial pigment spread inside, band form may be seen like *P. malariae*, but their cytoplasm is more irregular, and multiple parasites infecting 1 single red blood cell can be seen like in *P. falciparum*. examples of different parasite stages in the same thin blood slide image. In the first slide image, *P. falciparum* trophozoites and gametocytes can be seen together with white blood cells. The latter are larger and have a pronounced nucleus compared with the many red blood cells in the image. In the second image, *P. falciparum* ring stages are together with schizonts. In addition, other objects such as parasite outside cells and staining noise are visible in both images. Staining noise in particular can be confused with parasites by an unexperienced microscopist.

MALARIA DIAGNOSIS

Malaria is a curable disease, with drugs available for treatment, including drugs that can help prevent malaria infections in travelers to malaria-prone regions. However, there exists no effective vaccine against malaria yet, although this is an area of active research and field studies. Once infected, malaria is a rapidly progressing disease,

with a serious risk of developing into severe and cerebral malaria with neurologic symptoms for *P. falciparum* infections. Therefore, a timely diagnosis of malaria is very important. Although malaria can be diagnosed in many different ways, there is room for improvement for current malaria diagnostic tests including reducing cost, increasing specificity, and improving ease of use. Because automated malaria diagnosis for resource-poor settings is the main topic of this survey, we have devoted 2 subsections to light microscopy and RDTs, which are by far the 2 most heavily used diagnostic means in these areas. We also briefly discuss the other options for malaria diagnosis, although they are arguably less suited for the conditions in remote malaria regions. For more information about malaria diagnosis, we refer readers to the surveys and the following references. Detecting the presence of parasites is the key to malaria diagnosis. In addition, identifying the parasite species and presence of potentially mixed infections is important, as well as the observation of the stage development of *P. falciparum* parasites in relation to the severity of the disease. Counting parasites for determining the level of parasitemia is not only important for identifying an infection and measuring its severity, it also allows monitoring patients by measuring drug efficacy and potential drug resistance. Light microscopy. The current gold-standard method for malaria diagnosis in the field is light microscopy of blood films, which is the main focus of this article. Although other forms of diagnosis exist and have become popular in recent years, in particular RDTs, microscopy remains the most popular diagnostic tool, especially in resource-poor settings. With microscopy, all parasite species can be detected. It allows computing the level of parasitemia, clearing a patient after a successful treatment, and monitoring drug resistance. Furthermore, it is less expensive than other methods and widely available. However, its biggest disadvantages are the extensive training required for a microscopist to become a proficient malaria slide reader, the high cost of training and employing, maintaining skills, and the large component of manual work involved. To diagnose malaria under a microscope, a drop of the patient's blood is applied to a glass slide, which is then immersed in a staining solution to make parasites more easily visible under a conventional light microscope, usually with a 100× oil objective. Two different types of blood smears are typically prepared for malaria diagnosis: thick and thin smears. A thick smear is used to detect the presence of parasites in a drop of blood. Thick smears allow a more efficient detection of parasites than thin smears, with an 11 times higher sensitivity. On the other hand, thin smears, which are the result of spreading the drop of blood across the glass slide, have other advantages. They allow the examiner to identify malaria species and recognize parasite stages more easily. The actual microscopic examination of a single blood slide, including quantitative parasite detection and species identification, takes a trained microscopist 15–30 minutes. Considering that hundreds of thousands of blood slides are manually inspected for malaria every year, this amounts to a huge economic effort required for malaria diagnosis. Rapid diagnostic tests. The main advantage of microscopic malaria diagnosis lies in its low direct cost, which gives it a distinct advantage in resource-poor settings. Other existing diagnostic methods, and any new method, have to prove that they can provide the same ease of use and price point as microscopy given the limited financial resources typically available in malaria-prone regions. Arguably the only and main competitor in this sense are RDTs. They detect evidence of malaria parasites (antigens) and take about 10–15 minutes to process. Their detection sensitivity is lower but comparable with manual microscopy, and they do not require any special equipment and require only minimal training. Although RDTs are currently more expensive than microscopy in high-

burden areas, a valid question is whether these tests can replace microscopy in the near future. At the time of this writing, according to WHO, more countries use microscopy more than they use RDTs. RDTs are used more in rural areas where microscopy is not available. About 47% of malaria tests in malaria endemic countries worldwide were made by RDT. The use of RDTs, however, does not eliminate the need for malaria microscopy. A major disadvantage is that RDTs do not provide quantification of the results. Therefore, at this point in time, microscopy and RDTs are more complementing each other than one replacing the other. Other tests. Several methods for diagnosing malaria are available. Important criteria are cost per test, sensitivity and specificity of the method, time per test, and the required skill level of the user. Furthermore, quantification of the number of infected red blood cells is important as a prognostic indicator.

- Polymerase chain reaction (PCR). A molecular method called PCR has shown higher sensitivity and specificity than conventional microscopic examination of stained peripheral blood smears. In fact, it is considered the most accurate among all tests. It can detect very low parasite concentrations in the blood and can differentiate species. However, PCR is a complex highcost technology that takes many hours to process by trained staff. According to Tangpukdee et al., PCR is not routinely implemented in developing countries because of the complexity of the testing and the lack of resources to perform these tests adequately and routinely. Quality control and equipment maintenance are also essential for the PCR technique, so that it may not be suitable for malaria diagnosis in remote rural areas or even in routine clinical diagnostic settings.

- Fluorescent microscopy. Quantitative buffy coat is a laboratory test to detect infection with malaria or other blood parasites, using fluorescent microscopy. A fluorescent dye makes parasites visible under ultraviolet light. According to Adeoye and Nga, this test is more sensitive than the conventional thick smear. Nowadays, portable fluorescent microscopes with fluorescent reagent to label parasites, are available commercially. Although the quantitative buffy coat technique is simple, reliable, and user friendly, it requires specialized instrumentation, is more costly than conventional light microscopy, and is poor at determining species and numbers of parasites.

- Flow cytometry. This is a laser-based cell counting and detection method that allows to profile thousands of cells per second. Although flow cytometry offers automated parasitemia counts, this is offset by a rather low sensitivity. Flow cytometry is less suitable as a diagnostic technique in the field, when a direct answer is required for treatment decisions. However, in developed countries, it can be applied in the clinical setting for accurate counting of parasite numbers, for instance in the follow-up of drug treatment.

STAINING METHODS

More than 100 years ago, Giemsa stain (1902) was applied for the first time for the diagnosis of malaria. Since then, it received increased attention. Because of its low cost, its high sensitivity, and specificity, it is currently widely used in microscopical malaria examinations. However, Giemsa staining requires multiple reagents, experienced personnel, and is labor-intensive and time-consuming (it typically requires at least 45 minutes to stain a slide). Other stains have been used, too, like Field stain that significantly reduces the staining time, although it requires drying of samples before and during staining. However there are also disadvantages with Field's stain, specially

in under-resourced health centers in which the stain might be used. Poor blood preparations often result in the generation of artifacts commonly mistaken for malaria parasites, such as bacteria, fungi, stain precipitation, dirt, and cell debris. These can frequently cause false-positive readings. Another stain is Leishman's stain (1901), which has a high sensitivity, is cheap, and relatively easy to perform. Among the other stains being used is, for example, the Wright-Giemsa stain, which is a combination of Wright and Giemsa stain, and where the former facilitates the differentiation of blood cell types. In 1970s, Sodeman et al. investigated the effect of fluorochrome staining in identifying the malaria parasites at low-level infection. It has been shown that fluorochrome staining is more sensitive and less time consuming than Romanowsky and Giemsa staining methods but requires considerable practice and training, and suffers from artifacts including photobleaching and phototoxicity. Moreover, fluorescence microscopes are more expensive than standard light microscopes, which is a factor in tropical resourcepoor regions where malaria is endemic. the blood smear types and staining techniques used for the approaches published in the literature. Clearly, the vast majority of publications has been for thin smears. Certainly, 1 reason for this lies in the fact that thin smears allow to determine the parasite species and stages more easily, in addition to the parasitemia. So, in some sense, thin smears are more versatile and contain more information. Another important reason is probably that the presence of red blood cells gives the problem of parasite detection more structure, and makes the problem easier to a certain degree, as parasites need to be detected only inside cells. For thick films, parasite detection may be harder because of noise and staining artifacts that can lead to false positives. Nevertheless, because of the importance of thick smears for practical malaria diagnosis, it is very likely that more approaches for thick films will be implemented in the future. However, if convincing optical hardware solutions are found to scan multiple fields in thin smears and achieve a sensitivity comparable with thick smears, then this may be a moot point. also shows that the majority of approaches, for both thin and thick smears, have adopted the most popular stain in practice, Giemsa. Although stains like Leishman provide very good results for malaria parasites, Giemsa stain has proved to be the best all-round stain for the routine diagnosis of malaria. It has the disadvantage of being relatively expensive, but this is outweighed by its stability over time and its consistent staining quality over a wide range of temperatures.

AUTOMATED DIAGNOSIS OF MALARIA

This section provides the core information of our survey, namely a compilation of references that should cover the vast majority of articles ever published on automated microscopy for malaria diagnosis, with the bulk of the articles published in the last 10 years. The work that has been done in this area is quite diverse. Nevertheless, a system for automated cell microscopy usually implements a sequence of key processing steps that can serve as a guideline. Therefore, each of the following sub sections will focus on 1 specific aspect of the processing pipeline. The first step is usually the acquisition of digital images of blood smears, which largely depends on the equipment and materials being use. The Image acquisition section breaks down the different approaches for the different types of microscopy, blood slides (thin or thick), and staining. Following image acquisition, most systems perform one or several preprocessing methods to remove noise and to normalize lighting and color variations inherent in the image acquisition and staining process. The Pre

processing section sorts the publications according to the preprocessing methods implemented. The next step usually involves the detection and segmentation (outlining) of individual blood cells and maybe other objects that can be visible in a blood slide image, such as parasites or platelets. The section titled Red blood cell detection and segmentation gives an overview of all the segmentation methods that have been used for microscopic malaria diagnosis. For most articles, cell segmentation is followed by the computation of a set of features, which describe the visual appearance of the segmented objects in a mathematical succinct way. The section titled Feature extraction and selection presents the different features and potential feature selection strategies that can be found in the literature. In the last step, a mathematical discrimination method that classifies the segmented objects into different classes based on the computed features is implemented. For example, labeling each red blood cell as either infected or uninfected is a key classification task performed in this step, which then allows to compute the parasitemia. The section titled Parasite identification and labeling lists all the classification methods used in the literature for malaria diagnosis. Later in the article, in the section titled Deep Learning, we will present references for the latest classification trend, deep learning, which skips the feature computation step and sometimes even the segmentation step. Furthermore, in the section titled Mobile Smartphones for Malaria Diagnosis, we will discuss how smartphones can be used for microscopic malaria diagnosis and list the systems that have already been implemented and published. Image acquisition. lists all published systems according to the type of microscopy used. Because light microscopy is the most common form of malaria diagnosis in resource-poor settings, where automation will also have the largest impact on health care and economy, it is not surprising that most authors implemented systems for standard microscopy. We have also added all other imaging techniques that we found in the literature and for which automated systems have been developed. For more detailed information about these approaches, we refer to the references listed in the table and the reference list at the end of this article. Preprocessing. lists all preprocessing approaches that have been applied to automatic analysis of digital blood slide images. Preprocessing is mainly applied to improve the quality of the image and to reduce variations in the images that would unnecessarily complicate the subsequent processing steps. Three key objectives can be identified: noise removal, contrast improvement, illumination and staining correction. For noise removal, the most popular approaches have been well-established filters, such as mean and median filters, or Gaussian low-pass filtering. In addition, applying morphologic operations is very popular. For contrast improvement, contrast stretching techniques and histogram equalization in particular, have been the most popular approaches. For illumination and staining variations, color normalization techniques have been applied, including the popular use of grayscale colors. Red blood cell detection and segmentation. shows the different segmentation techniques applied to thin smears. The vast majority of these techniques are thresholding techniques, such as Otsu thresholding in combination with morphologic operations. However, these techniques may not be dominating because of their superior performance compared with other methods, but rather because of their relative simplicity. Other methods include Hough transform, which makes assumptions about the blood cell shape, and unsupervised k-means pixel clustering. Cell segmentation needs to be accurate to compute the correct parasitemia. However, touching cells in particular complicate the identification and segmentation of individual cells. For this problem, methods like watershed and active contours have been applied. shows the different segmentation

techniques in the literature for thick smears. The segmentation situation for thick smears is different in that white blood cells and parasites need to be segmented. However, white blood cells are bigger than red blood cells and have more texture, which makes their segmentation much easier. Furthermore, white blood cells just need to be identified and not to be processed or classified further. In addition, parasites are very small and their reliable identification is most important. Therefore, the detection of these objects is practically more important than their segmentation, which may explain again the dominance of thresholding techniques and morphologic operations. Feature extraction and selection. lists the different features used in the literature to describe the appearance of red blood cells, infected and uninfected, in thin smears. Obviously, because parasites have been stained, color features are most natural and indeed used by many articles. In addition, several texture and morphologic features have been used to describe the inside of red blood cells. The idea is that in case of infected cells, these features can pick up the typical appearance of ring structures with visible cytoplasm and other unique parasite characteristics. Generally speaking, most of the features used are tried and trusted features that have already been applied in other, often nonmedical, application domains. For example, Haralick's texture features, local binary patterns, co-occurrence matrices, histogram of gradients, and many others have been successfully used across a wide range of applications. This also includes morphologic shape features and moments. Most notably, here is the use of different color spaces, which leads to sets of more malaria-specific features, depending on the color space used. Although most articles remain in the standard RGB color space, we think that there is a perfectly good reason to use a different color space better suited to extract the typical staining colors, which often range from a blue or purple to brownish shade. The HSV color space is favored by many articles, and several other articles use the green channel of RGB to extract staining-related color information in gray scale. the features used for thick smears. Because of the smaller number of publications for thick smears, a smaller number of features has been experimented with in the literature. Nevertheless, authors have used similar, if not identical, features compared with the ones used for thin smears, experimenting with established features as well as different color spaces. Some articles compute a large set of many different features, and then for practicality reasons cut down on these features by selecting the most discriminative feature subset using feature selection strategies. Specifically, the feature selection techniques used to reduce feature dimensionality include principal component analysis, F-statistic, 1-way analysis of variance, information gain, and support vector machine-based recursive feature elimination. However, such classical approaches to feature computation and selection run the serious danger of being superseded soon by techniques not relying on handcrafted features, such as deep learning in particular, which we will discuss in the section titled Deep Learning. Parasite identification and labeling. lists all classification methods that have been used for either discriminating between infected and uninfected red blood cells in thin smears or identifying parasites in thick smears. Virtually all classification methods popular in the last decade have been applied to malaria diagnosis, ranging from decision trees and basic artificial neural networks over support vector machines to random tree classifiers. Very few articles have developed classification technologies specifically for cell discrimination or parasite detection. Most of the malaria-specific domain knowledge lies in the interplay of segmentation, features, and classification. Comparing the performance of the published systems is very hard. The systems have been evaluated on blood slides from entirely different origins with largely varying

parameters for image acquisition and slide preparation. Very often the evaluation set is too small or too limited to allow making a statement about the general system performance. Currently, there exists no publicly available image benchmark set, small or large, which could be used for fair comparisons of systems. Therefore, although many articles are reporting quite high performance numbers in terms of accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve, we prefer not to compare these numbers in this survey article. We can observe a trade-off between the processing pipeline's run-time performance and its accuracy. Typically, as the accuracy of a technique increases, its computational complexity increases all the same. For example, sophisticated level-set methods for cell segmentation perform better than Otsu thresholding but also require a longer runtime. Furthermore, feature computation can affect system efficiency. Some articles therefore apply feature selection methods to reduce feature dimensionality and remove nondiscriminative features, which can improve both accuracy and efficiency. Finally, the runtime of cell classification depends on the classification architecture used. For example, a support vector machine's classification is much faster than the classification by a deep neural network. Although many articles do not report runtimes for their systems, we think that most of the cited systems will perform their task many times faster than a microscopist, or at least will perform faster than a human after a little optimization of their implementation. We have also found 2 articles in which the authors developed dedicated hardware devices with motorized stage units to increase throughput. In combination with software, this will fully automate the slide screening process so that a microscopist does not need to move the microscope dish to take an image of the next field. This will also result in a higher throughput that can increase the sensitivity of the system by allowing to inspect more fields in the same time. To improve system accuracy, there seems to be a trend to follow the mainstream classification method at the time of publication to take advantage of the latest classification architecture and performance improvements it brings. Consequently, we are now seeing the first deep learning articles entering the scene, as listed in the next section.

DEEP LEARNING

Deep learning is the latest trend in machine learning, which has already boosted the performance in many non medical areas. Deep learning can be seen as an extension of the well-known multilayer neural network classifiers trained with back-propagation, except that many more layers are used. There are also different kind of layers that are used in typical successions. Deep learning typically requires large training sets. This is the reason why medical applications have been among the last applications to adopt deep learning, as annotated training images are significantly harder to obtain because of expert knowledge requirements and privacy concerns. The first article to apply deep learning to malaria diagnosis is by Liang et al., who use a convolutional neural network to discriminate between infected and uninfected cells in thin blood smears, after applying a conventional level-set cell segmentation approach. This is an ideal application for deep learning because images of segmented red blood cells are a natural input for a convolutional neural network. Deep learning does not require the design of handcrafted features, which is one of its biggest advantages. Other authors who have applied deep learning to cell segmentation are Dong et al. and Gopakumar et al., who used convolutional neural networks, Bibin et al., who used deep belief networks, and recently Hung et al. who presented an end-to-end framework

using faster Region-based Convolutional Neural Network. Because deep learning is the overarching machine learning technique nowadays, we can expect many more publications to appear soon for cell classification, cell staging, cell segmentation, and other sub-problems in automated malaria diagnosis.

MOBILE SMARTPHONES FOR MALARIA DIAGNOSIS

The ideal hardware solution for microscopic malaria diagnosis in resource-poor settings would be a small portable slide reader into which a blood slide could be inserted and which would then output the parasitemia. Although modern technology is heading this way, we are still far from having a field-usable device. In particular, the relatively high optical magnification needed (up to 1000 $\times$) for malaria diagnosis in combination with oil immersion is a major miniaturization obstacle, unless alternatives are found. The next best solution are small camera-equipped computing devices, such as smartphones, which can be attached to a magnifying device and can then compute the parasitemia automatically, using image analysis and machine learning. Modern smartphones have become powerful computing devices and their cameras provide sufficient resolution for malaria diagnosis. Moreover, Android phones have become relatively cheap and are often already in the possession of health-care workers, even in resource-poor settings. Although cellular network connectivity can help with the information exchange between field workers and hospital, it is not immediately needed for malaria diagnosis and the actual cell counting. Small magnifying devices that can be attached to a smartphone's camera, allowing true optical magnification compared with mere digital zooming, are commercially available. However, from the authors' experience, these devices are still lacking in the image quality provided. Therefore, a more practical approach is to simply attach the smartphone to the eyepiece of a regular microscope with an adapter so that blood slide pictures can be taken with the smartphone's camera. A few experimental set-ups along these lines have been reported in the literature. In Breslauer et al. built a mobile phone-mounted light microscope and demonstrated its potential for clinical use by imaging *P. falciparum*-infected and sickle red blood cells in brightfield and *Mycobacterium tuberculosis*-infected sputum samples in fluorescence with LED excitation. In all cases, resolution exceeded that necessary to detect blood cell and microorganism morphology. For tuberculosis samples, they took advantage of the digitized images to demonstrate automated bacillus counting via image analysis software. In Pirnstill and Cote present a cost-effective, optical cellphone-based transmission polarized light microscope system for imaging the malaria pigment known as hemozoin, which is a disposal product of the parasite's blood digestion. It can be difficult to determine the presence of the pigment from background and other artifacts, even for skilled microscopy technicians. The pigment is much easier to observe using polarized light microscopy. However, implementation of polarized light microscopy lacks widespread adoption because the existing commercial devices have complicated designs, require sophisticated maintenance, tend to be bulky, can be expensive, and would require re-training for existing microscopy technicians. The cellphone-based polarimetric microscopy design presented by Pirnstill and Cote shows the potential to have both the resolution and the specificity to detect malaria in a low-cost, easy-to-use, modular platform. Rosado et al. presented an image processing and analysis methodology using supervised classification to assess the presence of *P. falciparum* trophozoites and white blood cells in Giemsa-stained thick blood smears. Using a support vector

machine and a mix of geometric, color, and texture features, their automatic detection of trophozoites achieved a sensitivity of 80.5% and a specificity of 93.8%, whereas their white blood cell detection achieved 98.2% sensitivity and 72.1% specificity. In Quinn et al. presented their 3-dimensional printable design of an adapter to attach a smartphone to a microscope, although all images for their experiments were taken with a dedicated microscope camera, which offered a higher pixel resolution than their smartphone camera. They presented a workflow for automated analysis of thick blood smears, which involved the computation of morphologic and moment features and an ensemble tree classifier trained on these features to discriminate between abnormal patches containing parasites and normal patches. The performance they reported was 97% area under the receiver operating characteristic curve. Skandarajah et al. built a custom mobile phone microscope that is compatible with phones from multiple manufacturers. They demonstrated that quantitative microscopy with micron-scale spatial resolution can be carried out with multiple phones and that image linearity, distortion, and color can be corrected as needed. Specifically, they showed that phones with greater than 5 megapixel cameras are capable of nearly diffraction limited resolution over a broad range of magnifications, including those relevant for single cell imaging. Furthermore, they found that automatic focus, exposure, and color gain standard on mobile phones can degrade image resolution and reduce accuracy of color capture if uncorrected, and they devise procedures to avoid these barriers to quantitative imaging. Dallet et al. describe a mobile application platform for Android phones that can diagnose malaria from Giemsa-stained thin blood film images. The main imaging component consists of elaborate morphologic operations that can detect red and white blood cells, and identify parasites in the infected cells. The application also recognizes the different life stages of parasites and calculates the level of parasitemia. The application takes less than 60 seconds to give a diagnosis, and has been tested and verified on several versions and types of Android mobile phones and tablets. The authors of this survey article have developed a smartphone application to compute parasitemia in Giemsa-stained thin blood film images. To segment individual red blood cells, we applied marker-controlled watershed to thin blood smears to efficiently detect and segment individual cells, separate touching cells, and meet the demand of real-time processing. In the cell detection step, we apply a multiscale Laplacian of Gaussian filter on the green channel of an RGB color slide image. The local extrema of the Laplacian of Gaussian response indicate the approximate centroids of the individual cells that will serve as the approximate centroids for the marker-controlled watershed segmentation step. The cell foreground mask is estimated using Otsu thresholding, and cell edges are extracted by computing the gradient magnitude over the minimum values of the green and blue channels. Then, in the segmentation step, we apply watershed transform on cell markers, foreground masks, and edge information to segment and separate touching cells. For cell classification, we follow a deep learning approach and use a convolutional neural network for discriminating infected from uninfected cells. Cesario et al. discuss mobile support for vectorborne diseases in areas where specialist health care is scarce. They focus on the image analysis and classification component of a system that aims to reduce the chance of misdiagnosing less common diseases as malaria and to assist health professionals. Their article largely describes work in progress toward the image analysis and classification component, but feedback from healthcare professionals has been generally positive. Herrera et al. tested the diagnostic performance of a device for automated interpretation of RDTs, which uses smartphone technology and image analysis software. he diagnostic

performance of the device was comparable with visual interpretation of RDTs, without significant differences for *P. falciparum* and *P. vivax*. Providing standardized automated interpretation of RDTs in remote areas, in addition to almost real-time reporting of cases and enabling quality control, would greatly benefit large-scale implementation of RDT-based malaria diagnostic programs. In similar work, Mudanyali et al. demonstrated a cellphone-based RDT reader platform that can work with various lateral flow immuno-chromatographic assays and similar tests. Their compact and cost-effective digital RDT reader attaches to the existing camera unit of a cellphone, where RDTs can be inserted to be imaged. Captured raw images of these RDTs are then digitally processed through a software application running on the cellphone for validation of the RDT and for automated reading of its diagnostic result. In addition, this smart RDT reader platform running on cellphones provides real-time spatio-temporal statistics for the prevalence of various infectious diseases, which allows tracking epidemics.

CONCLUSION

We wrote this survey article on image analysis and machine learning methods to give an update on the latest development in automated malaria diagnosis with image analysis and machine learning. This is a very dynamic area of research that has seen an extensive number of publications in the last decade. However, with the advent of new deep learning approaches, which have already left a deep impression, the research is seeing a new exciting development that is nothing short of a revolution. So far, only a few articles have been published, but it is already evident that this will be the dominating technique in the foreseeable future. This will render many of the former classification approaches dispensable. Moreover, because deep learning takes the difficult task of designing features for classification from the user, many of the handcrafted features used so far may become useless. In addition, because deep learning can be used not only for cell classification but also for cell segmentation, many of the cell segmentation approaches presented so far could become outdated very soon. Even the preprocessing techniques, which play an important role, are not safe from this development. One way of thinking is that neural networks can learn how to process different staining and lighting variations if only enough training data are being presented to the network. Given the recent developments and future possibilities, there is in fact a good chance that most of the articles referenced in our and other surveys will become a mere historical side note very soon, describing the state of the art before the advent of deep learning. All of the deep learning articles published so far have concentrated on thin blood smears, but it is very likely that we will see articles for thick films very soon. Given the wide acceptance of deep learning, the importance of large annotated data image repositories for training is now widely understood, leading to a great support of data acquisition efforts. This will likely lead to larger test suites on patient level, allowing for more standardized evaluations and extensive field testing. Given these developments, automated microscopy is very much in the race toward a cheap, simple, and reliable method for diagnosing malaria.

2.3. MALARIA DETECTION USING MACHINE LEARNING WITH K NEAREST NEIGHBOUR ALGORITHM

Abstract:

To detect the presence of Malaria parasites in Human Blood Cells and to determine how much it is affected. We have developed a machine learning method that can detect malaria parasites in thick blood smear images. Our method consists of two processing steps. To start with, we apply a force based Iterative Global Minimum Screening (IGMS), which plays out a quick screening of a thick smear picture to discover parasite up-and-comers. Then, a customized K-Nearest Neighbor (K-NN) classifies each candidate as either parasite or background. Malaria - a parasitic disease health problem which leads to millions of deaths especially in remote villages. This disease arises due to damaging of red blood cells. This paper presents a survey on detection / prediction of malaria disease using various machine learning techniques, Image Processing techniques and various clinical methods like rapid test, Nested PCR etc. In our observation, we found that machine learning techniques have wider applicability for critical diagnosis of malaria which in turn helps the clinicians for diagnosing the disease.

Introduction:

An overview of detecting malaria disease based on data classification technologies, screening, pre-existing systems and their drawbacks, motivation behind the project, its objective and limitations, the functional and the non- functional requirements upon which the project is developed. This paper highlights the prediction of malaria diseases and the different factors responsible for it. It briefly describes the different data classification techniques in this work used for comparison.

Scope:

Malaria diseases depends on a lot of variables like hypertension, diabetes, exercise schedules, chest pain etc. This risk factors may not be scaled equally to every individual patient. By the help of affected human grey scale image, the reports are given to the doctors for immediate actions to be taken. Medical databases are usually filled by irrelevant and redundant features which increase the dimension of database or lead them to curse of dimensionality. It affects the accuracy, computational cost and speed of the learning process. The accuracy and sensitivity have particular importance in the detection and prediction of diseases. Its positive feedback can predispose until the doctor by its analysis to speed up the process of diagnosis and prognosis.

Parasite PreSelection Using Iterative Global Minimum Screening (IGMS):

IGMS generates RGB parasite candidates by localizing the minimum intensity values in a grayscale image. If only one pixel is localized, a circular region centered at this pixel location with a pre-defined radius of 22 pixels (average parasite radius) is

cropped from the original RGB image and is selected as a parasite candidate. If more than one pixel is localized, a new parasite candidate centered at the i th pixel is added when all the distances between the i th pixel and previously selected pixels are larger than 22. Once a parasite candidate is selected, the intensity values inside this region of the grayscale image will be replaced by zeros to guarantee the convergence of the IGMS method. The screening stage stops when the number of parasite candidates reaches a given number. Experiments on our dataset of 150 patients show that we can achieve a sensitivity above 97% on patch level, image level, and patient level when using this number. Each parasite candidate is a $44 \times 44 \times 3$ RGB patch image, with pixels having a distance greater than 22 to the center set to zero.

K-Nearest Neighbor Algorithm:

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until function evaluation. Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k -NN classification) or the object property value (for k -NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

Malaria Disease Classifier (K-NN):

Once the parasite candidates are extracted, we use a K-NN model to classify them either as true parasites or background. In this work, we customize a K-NN model consisting of seven convolutional layers, three max-pooling layers, three fully connected layers, and a softmax layer as shown in batch normalization layer is used after every Convolution layer to allow a higher learning rate and to be less sensitive to the initialization parameters, followed by a rectified linear unit (ReLU) as the activation function. Max-pooling layers are introduced after every two successive convolutional layers to select feature subsets. The last convolutional feature map is connected to three fully connected layers with 512, 50, and 2 hidden units, respectively. Between the three fully connected layers, two dropout layers with a dropout ratio of 0.5 are applied to reduce model over fitting. The network is derived from VGG19 by selecting the first six convolutional layers and three corresponding max-pooling layers from the VGG19 architecture to stop the feature maps, followed directly by the fully connected and dropout layers. This shorter network structure provides similar performance while being faster and requiring less memory, which is important for smartphone applications. The output of the K-NN model is a score vector, which gives the probabilities of the input image patch being either a parasite or background. We can obtain a larger or smaller number of predicted parasites by applying an adaptive probability threshold to the score vector. Compared with pre-trained networks such as VGG, GoogLeNet, ResNe-50.

K-Nearest Neighbor Algorithm Steps:

1. Load the data
2. Initialize K to your chosen number of neighbors
3. For each example in the data
 - 3.1 Calculate the distance between the query example and the current example from the data.
 - 3.2 Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels

2-D Feature Extraction:

In this work, the front-end CNN is same as the first ten layers of VGG-16 with three pooling layers, considering the tradeoff between accuracy and the resource overhead. VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU's. There are 13 convolutional layers, 5 Max Pooling layers and 3 dense layers which sums up to 21 layers but only 16 weight layers. Conv 1 has number of filters as 64 while Conv 2 has 128 filters, Conv 3 has 256 filters while Conv 4 and Conv 5 has 512 filters. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 .

Segmentation by dilated KNN:

The back-end CNN is a series of dilated convolutional layers producing density map. The goal of semantic segmentation is classifying each pixel of the input image into a given set of classes. The main challenge of this is to combine pixel level accuracy with multi scale contextual information. The previous state of the art models are based on the adaptations of convolutional neural networks designed for image classification. The idea of Dilated Convolution was motivated as it enlarge receptive field while maintaining resolution and also it come from the wavelet decomposition. It is also called "atrous convolution" and "hole algorithm". This module is modified version of adapted VGG-16 network for semantic segmentation by removing the last two pooling and striding layers. Dilated convolutions utilize specific kernels with sparsely aligned weights. Both of the kernel size and the interval of sparse weights expand exponentially with dilation factor. By increasing dilation factor, receptive field is also expanded exponentially by large kernel. Dilated Convolution (Basic or Large) can always improve the results and does not overlap with any other post- processing steps. A dilated convolution is essentially a generalization of the traditional 2D convolution that allows the operation to skip some inputs. This enables an increase in the size of the filter (i.e. the size of the receptive field) without losing resolution.

Output using K-NN Algorithm:

A good classification algorithm's accuracy must be high enough for better classification. Figure shows the accuracy of the classification results. K-NN enumerates the malaria disease diagnosis which consists of fully malaria related clinical data. Among the data record, K-NN provides better result with high accuracy. The attained accuracy shows the effectiveness of the MRF algorithm.

Conclusion:

In this paper it is an overview of using Iterative Global Minimum Screening (IGSM) as a feature extraction with K-NN classifiers. It was observed that the performance rate of the classifiers after using IGSM to reduce the dimension of data improved. Experiments are conducted on 500 data records. Results show that using IGSM encouragingly enhances classification performance on most of the classifiers. From the graphs it is interpreted that the K-NN not only increases the support and confidence but also the run time taken is reduced considerably. Further implementation of the K-Nearest Neighbor (K-NN) will be generating the results which will be analysed.

2.4. Malaria Detection Using Image Processing and Machine Learning

Abstract

Malaria is mosquito-borne blood disease caused by parasites of the genus *Plasmodium*. Conventional diagnostic tool for malaria is the examination of stained blood cell of patient in microscope. The blood to be tested is placed in a slide and is observed under a microscope to count the number of infected RBC. An expert technician is involved in the examination of the slide with intense visual and mental concentration. This is tiresome and time consuming process. In this paper, we construct a new image processing system for detection and quantification of plasmodium parasites in blood smear slide, later we develop Machine Learning algorithm to learn, detect and determine the types of infected cells according to its features.

INTRODUCTION

Malaria is a mosquito-borne life threatening disease caused by *Plasmodium* parasite. Globally, an estimated 3.2 billion people are at high risk (>1 in 1000 chance of getting malaria in a year). According to the report, there were 212 million new cases of malaria worldwide in 2015 (range 148–304 million). The WHO African Region accounted for most global cases of malaria (90%), followed by the South-East Asia Region (7%) and the Eastern Mediterranean Region (2%). Visual detection and recognition of *Plasmodium* in RBC is possible via chemical process. The staining process somewhat colorizes the RBCs but highlight *Plasmodium*, WBCs and platelets. The detection of *Plasmodium* requires detection of the stained objects. However, we need to analyzed stained objects further to determine if they are parasites or not to prevent false diagnosis. Several methods exist for malaria detection. Malaria parasite (MP) in blood sample can be identified by using image segmentation and feature extraction using minimum distance classifier. Based on Image Acquisition, Image

Preprocessing, Image Smoothing, Thresholding and Dilation image segmentation is done. Feature extraction uses two phases in architectural model:

- 1) Training Phase and
 - 2) Recognition Phase which helps to recognize the MP.
- In this work, we focus 1) automated detection and quantification of malaria detection,
2) strategy to determine infected image using machine learning
3) discuss to improve the predictive value for detection of infected cells .

METHODOLOGY

Our data consists of heterogeneous datasets, we have lot of variability in the images. So it might be easy to develop an algorithm that works for particular image but it is much different challenge to find an algorithm that works across the board of heterogeneous datasets. There are couple of approaches we can take we can try an algorithm that is robust to the heterogeneity or we can cached the data into more homogeneity state.

A. Data Collection

Digital image are the main data sources for this research. The images used in this research are taken from Vector-Borne Disease Research and Training Center (VBDRTC), Hetauda , Epidemiology and Disease Control Division, Teku, Kathmandu, Nepal. These images are of highly defined and magnified. Image of RBC's that are taken in thin smear are further processed and analyzed for the MP detection.

B. Purposed Architecture

Using malaria infected images which is of less noisy and devoid of artifact, this research can be done. The method used for the counting malaria infected RBCs Instead of finding how many cells are infected we will find what types of infection it is using machine learning. A machine learning algorithm takes examples of inputs and outputs associated with a task and produces a program that can automatically differentiate them. In this research we are not going to discuss about machine learning but we will show the architecture we used for this research.

1) RBC's Image Database:

Image acquisition is the first process in all the image processing steps. So, we will have RBC's image database where malaria infected RBC's and non-infected RBC's will be placed. Some of the images that are tested for MP Images that are taken first goes towards chemical process where pH will be maintain so that the MP will be visible clearly. is the input image which has two Malaria Parasite and some impurities. is the input image which does have Malaria Parasite but contains impurities. has one MP and shallow purple colored impurities. is the input image which contains one MP and impurities.

2) Global Thresholding

Segmentation is the process of separating an image into regions with respect to objects. We try to segment regions by identifying common properties. The simplest property that pixels in a region can share is intensity. So, a natural way to segment such regions is through thresholding, the separation of light and dark regions. Thresholding creates binary images from grey-level ones by turning all pixels below some threshold (T) to zero and all pixels above that threshold to one.

3) Enhancement

Image enhancement is the process of adjusting digital image so that the result is more suitable than the original one and will be helpful for further processing. During the research, input image will be enhanced after we threshold the image. After we threshold the we found the following result. From the result, we can see that the background and RBCs with parasite has been separated into white and black. For the better segmentation we will be doing complement of the above result.

4) Morphological Operations

Morphology is a tool for extracting image components that are useful in representation and description of region, shape. Morphological operations apply a structuring element to an input image, creating an output image of the same size. The value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors. Erosion and Dilation are fundamental steps for morphological processing. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. From it is noted that the pixels around the boundaries has been removed. Removing the pixel boundaries, noise can be reduced. Form it is noted that the pixels around the boundaries as been added. So, dilation helps in detecting the object clearly.

5) Segmentation

Segmentation subdivides an image into its constituent regions or objects. The level of detail to which the subdivision is carried depends on the problem being solved. Segmentation accuracy determines the eventual success or failure of computerized analysis procedures. For this reason, considerable care should be taken to improve the probability of accurate segmentation. Most of the segmentation algorithm is based on the basic properties of intensity values: discontinuity and similarity. During the research, it is found that two types of segmentation have to be implemented to get the result. Watershed segmentation and Color-based segmentation are used in the research.

a) Watershed Segmentation

Watershed segmentation is relatively a recent approach which tends to separate touching objects. This research applied Watershed segmentation process so that the overlapped RBCs will be separated and will be helpful for counting RBCs. In watershed segmentation, we assume 2D, grayscale image as a topological in which the location is given by the x,y image coordinates and the height at that location corresponds to the image intensity or grayscale value. A catchment basin defines that connected region for which any rainfall drains to the same low point or minimum. In terms of digital image, catchment basins consist of a group of connected pixels. Line which divides one catchment basin from another is called watershed ridges, watershed lines or simply watersheds. Rain which falls on the landscape will naturally drain downwards, under the action of gravity, to its nearest minimum point. We imagine the landscape being gradually flooded from below with the water entering through the local minima. As the water level increases, we construct dams which prevent the water from the catchment basins spreading into adjacent catchment basins. The construction process stops when the water level reaches the height of the highest peak. The dams built in this way are the watersheds which partition the landscape into distinct regions containing catchment basin. We can use original input image for the segmentation but for the suitable segmentation, gradient image is used because gradient magnitude is usually high along object edges and low elsewhere. So, the watershed ridges lie along the object edges. After the morphological operation, distance transform image is used. Distance transform is an operator that transforms

the values of pixels in an image according to their distance from the boundary of the object. Farther the pixel from the boundary, higher the value is returned. This will result in transformed image where the area near and within the boundary will be dark whereas the area near the center of the image will be white. Here the distance is calculated as Euclidean distance. shows the distance transform image after morphological operations. Euclidean distance i.e. the distance from the every pixel to the nearest nonzero is calculated and shows the distance transformed image of . Since the catchment basin in it is not shown properly, the distance transforms of the complement of and negate the distance is done and shown in . Then after, watershed segmentation is done which is shown in the . Noise and small-scale structures within the objects generally result in many local, small catchment basins. These spoil things and produce over-segmentation. To overcome this, minima that are too shallow are removed and do watershed in the result. Final watershed segmentation is

b) Color based segmentation

Color based segmentation is an approach to separate the defined color from other colored object. Blood samples are gone through the chemical process, and mostly the MP will be in the form of purple color. So, we segment the image with respect to RBCs cells .

The CIE has defined a system that classifies color according to the HVS (the human visual system). Using this system we can specify any color in terms of its CIE coordinates and hence be confident that a CIE defined color will match another with the same CIE definition. The CIE has measured the sensitivities of the three broad bands in the eye by matching spectral colors to specific mixtures of three colored lights. The spectral power distribution (SPD) of a color is cascaded with these sensitivity functions to produce three tri-stimulus values. These tri-stimulus values uniquely represent a color, however since the illuminant and lighting and viewing geometry will affect the measurements these are all carefully defined. The three CIE tri-stimulus values are the building blocks from which many color specifications are made. CIELab has no associated two dimensional chromaticity diagram and no correlate of saturation. CIELhs can therefore not CIE $L^*a^*b^*$ is an attempt to linearize the perceptibility of unit vector color differences. It is non-linear, and the conversions are still reversible. Coloring information is referred to the color of the white point of the system, subscript n. The non-linear relationships for L^* a^* and b^* are the same as for CIELUV and are intended to mimic the logarithmic response of the eye. Hue is an angle in four quadrants, and there is no saturation term in this system. When determining CIEL $^*a^*b^*$ values for CRT displayed colors it is usual to used the CRT's white point as the reference white.

6) Calculate and find perimeter of parasite and mean perimeter of RBC cells.

After we finished the most important section of our process, we now have to find the mean perimeter of the RBCs. From the research it is found that most of the parasites are larger than the mean perimeter of RBCs cells. With the help of matlab function regionprops we can easily detect the perimeter of object in the image. Regionprops helps to measure properties of image regions.

After finding the mean perimeter of RBCs cells, we will find the perimeter of each MP. Basically, malaria is transmitted by female mosquitoes of the genus Anopheles. This mosquito is itself a parasite, the females visiting humans for occasional meals of blood. During feeding, infected mosquitoes pass on the malaria parasite from their

salivary glands. The mosquito is described as vector. An immature form of plasmodium is injected into the blood of humans by the mosquito. This form disappears from the bloodstream as it enters various cells of the body, particularly the liver. Here it multiplies to produce large numbers of a form which can infect other liver cells. Each parasite in a red blood cell undergoes further division. The red blood cell bursts and released parasite can enter other red blood cells. As a result of this extensive division, millions of these parasites can be present in the blood. Some of the parasite transforms into male and female forms of the parasite. The image of malaria infecting human blood cell is shown : Malaria Transmission Cycle So, the RBC's are in the form of circular shape. When the RBC's are infected by MP, the shape of RBC's will be bursts and the shape of RBC's will be changed.

7) Compare whether the parasite is greater than mean value of RBC cells

This process comes under the feature extraction process. To detect infected and non-infected cell, we will compare the parasite's parameter with mean parameter of RBCs. If the perimeter of parasite is greater than the mean perimeter of RBCs cell, we will make a circle around the parasite. If the perimeter of the parasite is equal or less than the mean perimeter of RBCs cells we will do nothing.

8) Plot circle around infected RBC and calculate the infected cells. Parasite perimeter greater than the mean perimeter of the RBCs cells are plotted with red circle as shown . And then the number of infected RBCs will be count.

If the proposed methodology will able to find at least one RBC infected by the parasite then the research will be successful because the person is considered to be malaria infected if at least one out of multiple parasite can be detected. With the help of doctor and lab technician we can validate whether the red marked circle are infected RBCs or not.

RESULTS & DISCUSSION

During the research, more than 40 images have been tested and some of the testes images are listed below. We have tested for malaria tested non-infected RBCs cells in is the binarize output image after threshold. The mean perimeter of RBCs cells calculated is 318.0261 and the perimeter of parasite greater than mean perimeter of RBC is not found. So, the number of infected cells found is 0. In the, there are artifacts which have not been detected in final output which are true. Number of infected parasite is 1 which is not true because there is a MP adjacent to the red circle in Figure 25 which is not detected. Mean perimeter is 121.041. Greater than the mean perimeter of RBCs is detected parasite and mark circle around it. Same as the above process, has been tasted and gives the output in Figure 30. Mean perimeter of the RBCS is 119.7608. The output which gave 1 infected RBC which is true. As in input image 1 Figure 16, input image 4 has been tested. Mean perimeter of RBCs is 118.0082. Parasite greater than mean value is traced as circle. And finally found the number of infected cells 1 which is true. So, from the above result we have found that parasites present in the RBCS cells have been detected. Some errors are still found like not counting the entire infected parasite, impurities greater than mean will be counted as infected cells. Although there are errors, if we find at least one parasite in the RBCs than it is said that she or he is malaria infected. Remaining works Around 20/20 of the infected and non infected RBCs cells images have been tested till now. Validation and verification has to be done. Some errors have been found and need to detect in which stage MP has been detected and which stage of MP are not being detected. We have to find the error percentage.

CONCLUSION

We have presented an approach and developed an algorithm for detecting malaria, automated malaria detection and quantification of malaria infection. Also, we developed a strategy to train with machine learning, adaptable to detection of malaria with other types of parasite and also discuss to increase the predictive value with results.

Chapter 3

Existing system:

- Malaria is caused by protozoan parasites of the genus Plasmodium that are transmitted through the bites of infected female Anopheles mosquitoes and that infect the red blood cells
- falciparum is the most prevalent malaria parasite in sub-Saharan Africa, accounting for 99% of estimated malaria cases in 2016.
- Detecting the presence of parasites is the key to malaria diagnosis. In addition, identifying the parasite species and presence of potentially mixed infections is important, as well as the observation of the stage development of P. falciparum parasites in relation to the severity of the disease.
- Counting parasites for determining the level of parasitemia is not only important for identifying an infection and measuring its severity, it also allows monitoring patients by measuring drug efficacy and potential drug resistance.

Disadvantage:

- Using some of deep learning method.
- Image analysis is low efficiency.
- Result detection is not perfect.
- It did not detect the malaria by using of image.

Proposed system:

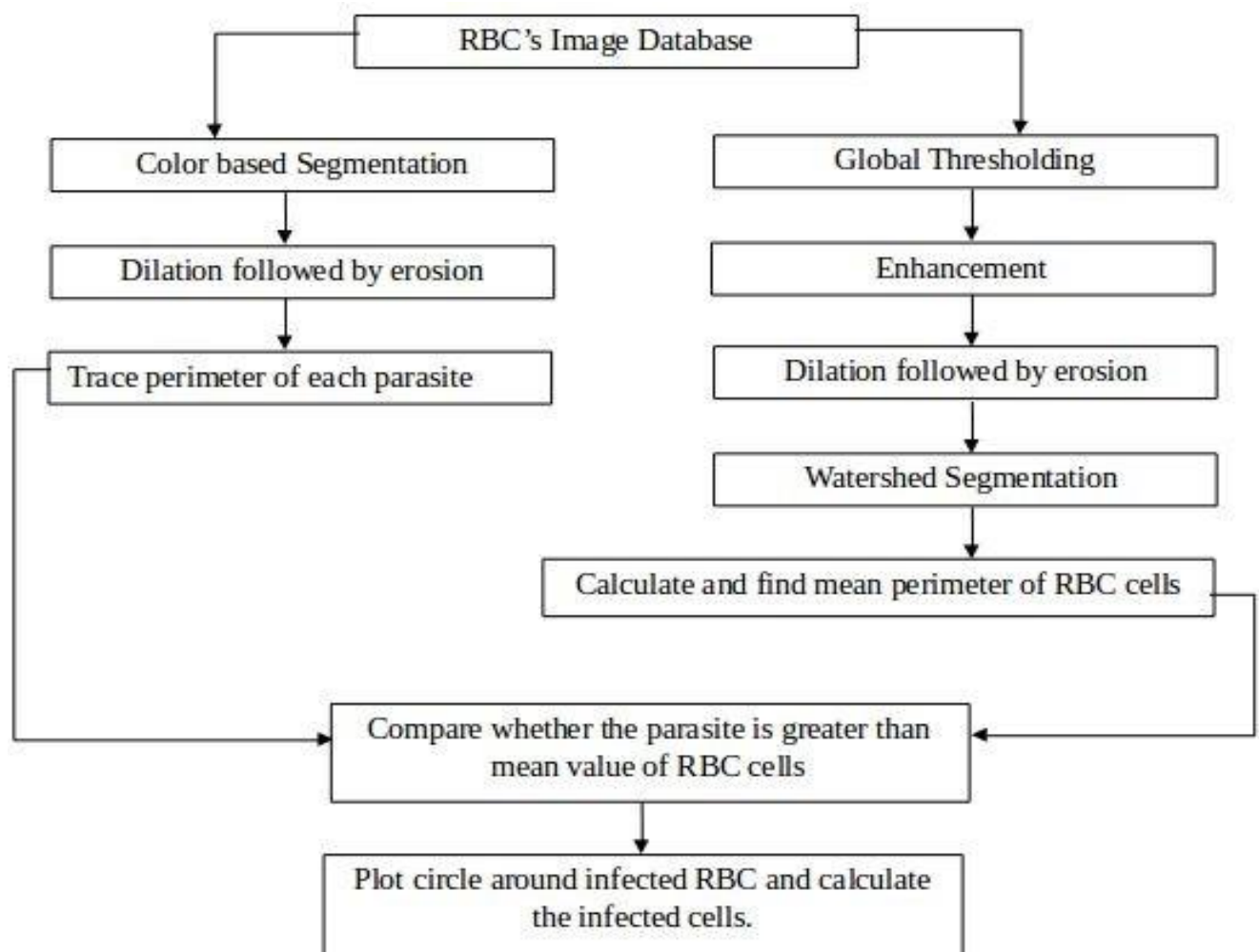
- Our data consists of heterogeneous datasets, we have lot of variability in the images.
- So it might be easy to develop an algorithm that works for particular image but it is much different challenge to find an algorithm that works across the board of heterogeneous dataset.
- There are couple of approaches we can take we can try an algorithm that is robust to the heterogeneity or we can cached the data into more homogeneity state.
- Digital image are the main data sources for this research. The images used in this research are taken from Vector-Borne Disease Research and Training Center (VBDRTC), Hetauda , Epidemiology and Disease Control Division, .

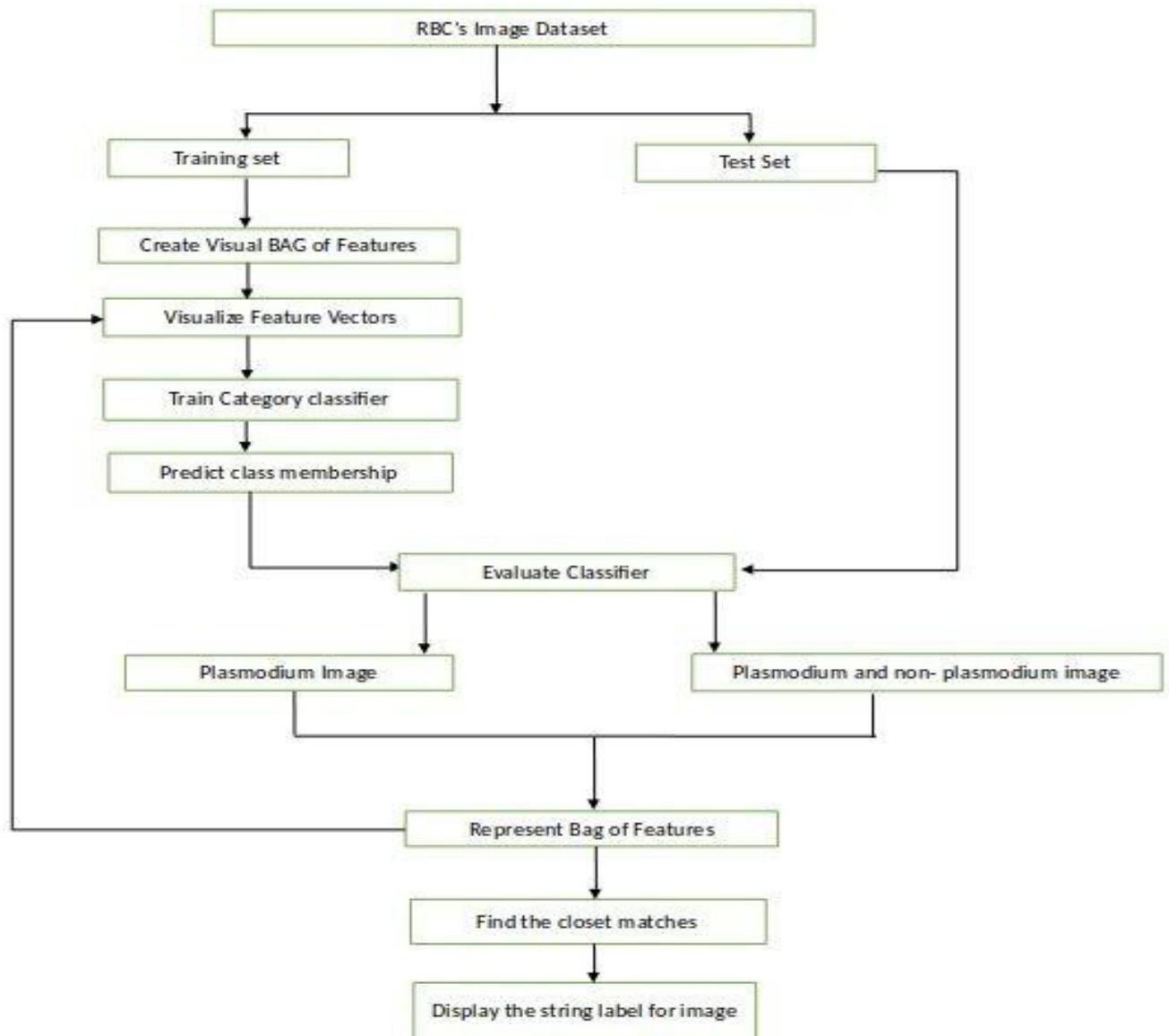
- These images are of highly defined and magnified. Image of RBC's that are taken in thin smear are further processed and analyzed for the MP detection.

Advantage:

- This system is combination of image processing and machine learning.
- Result is more efficiency.
- Image processing method used to process the image.
- And machine learning method used for detect the malaria.
- Using machine learning algorithm .

Block diagram:





Flow diagram:

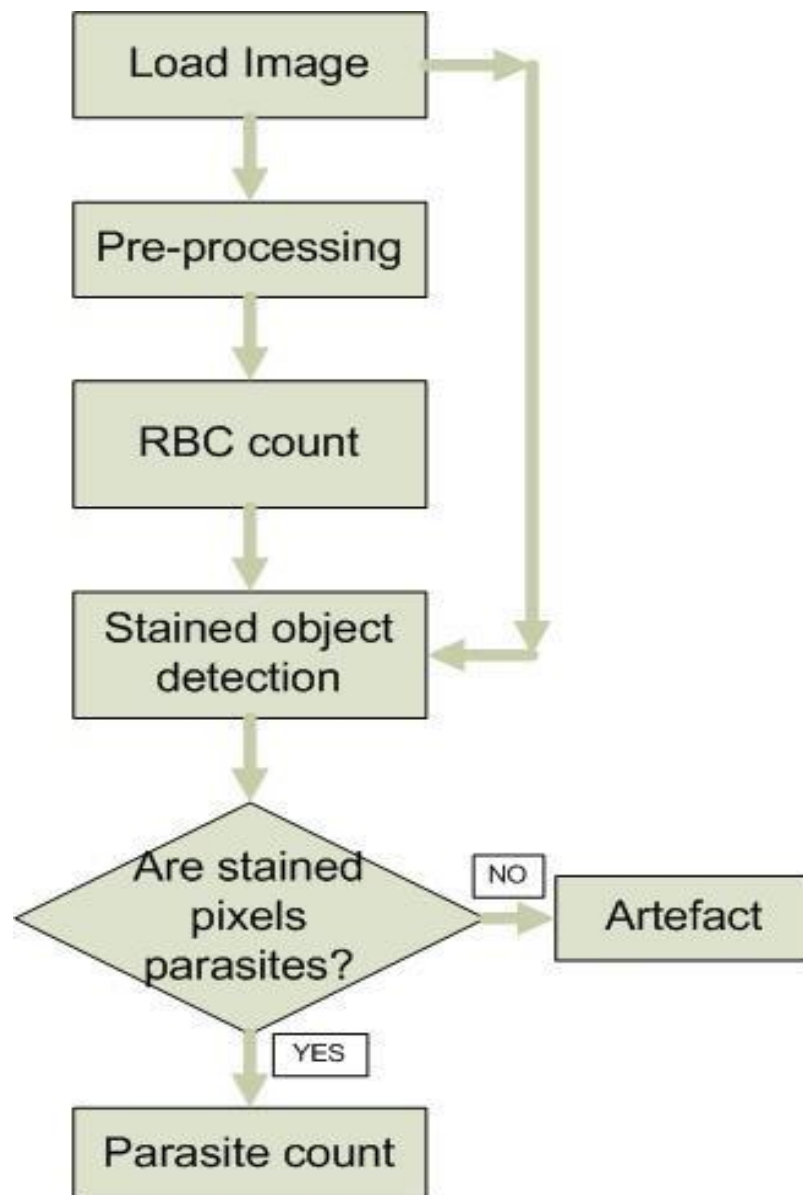
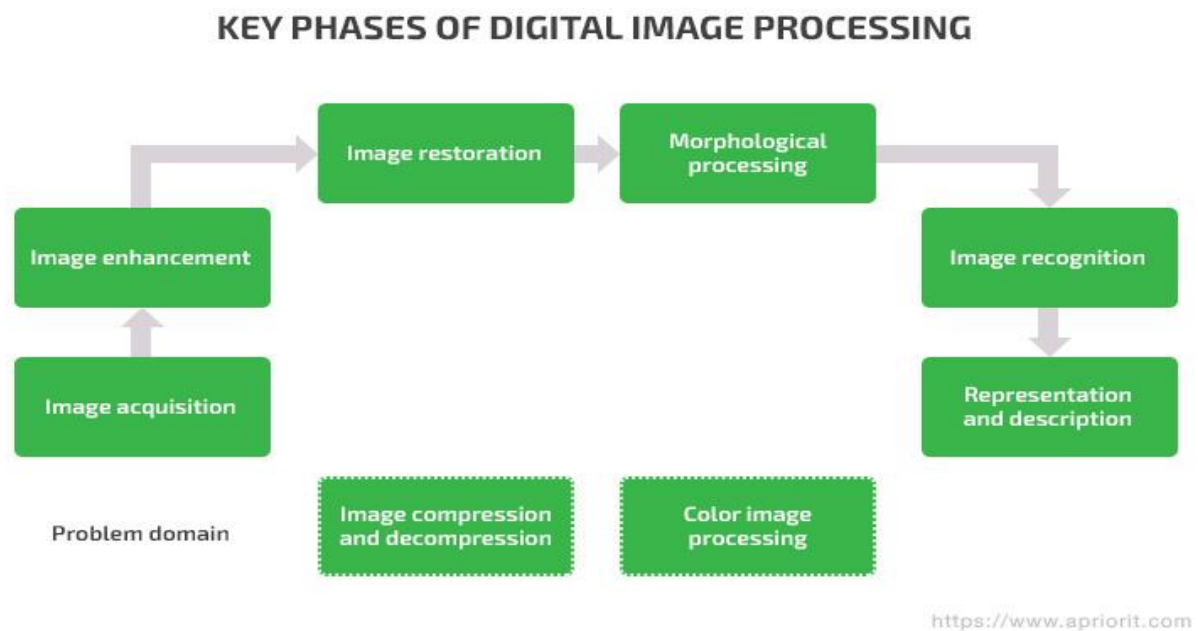


Image process:

- Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image.
- Digital Image Processing (DIP) is a software which is used to manipulate the digital images by the use of computer system. It is also used to enhance the images, to get some important information from it. For example: Adobe Photoshop, MATLAB, etc.

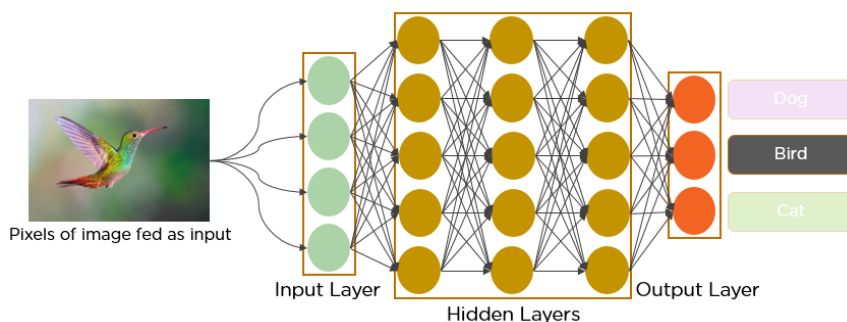
- Many advanced Image Processing methods leverage Machine Learning Models like Deep Neural Networks to transform images on a variety of tasks, like applying artistic filters, tuning an image for optimal quality, or enhancing specific image details to maximize quality for computer vision tasks.
- Machine Learning basically means that you're training the machine to do something (here, image processing) by providing set of training data's. Machine Learning have models/architectures, loss functions and several approaches that can be used to determine which would provide better image processing.



Convolutional Neural Networks (CNN)

Introduction:

In the past few decades, Deep Learning has proved to be a very powerful tool because of its ability to handle large amounts of data. The interest to use hidden layers has surpassed traditional techniques, especially in pattern recognition. One of the most popular deep neural networks is Convolutional Neural Networks.



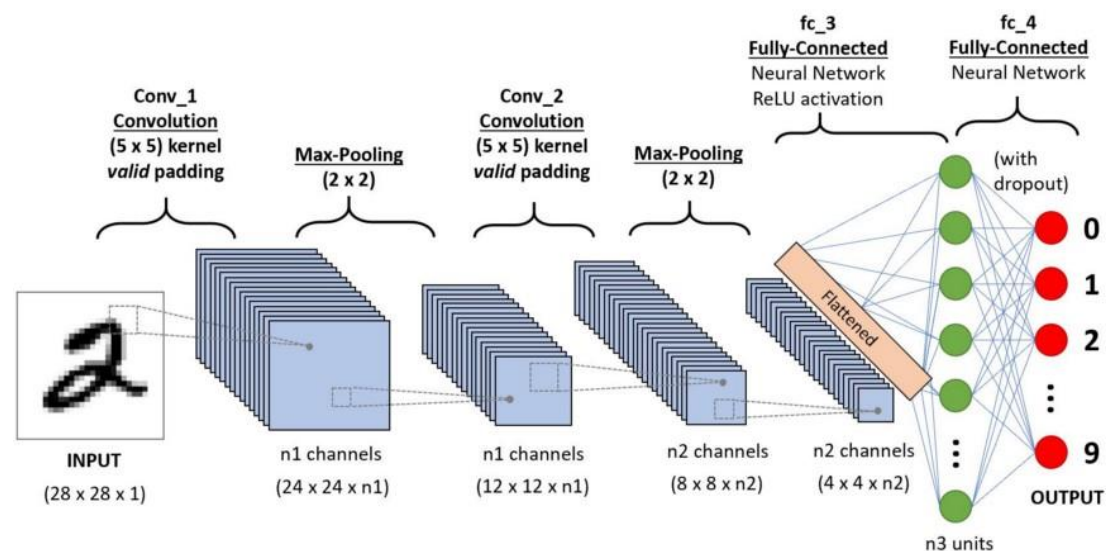
Since the 1950s, the early days of AI, researchers have struggled to make a system that can understand visual data. In the following years, this field came to be known as Computer Vision. In 2012, computer vision took a quantum leap when a group of researchers from the University of Toronto developed an AI model that surpassed the

best image recognition algorithms and that too by a large margin. The AI system, which became known as AlexNet (named after its main creator, Alex Krizhevsky), won the 2012 ImageNet computer vision contest with an amazing 85 percent accuracy. The runner-up scored a modest 74 percent on the test. At the heart of AlexNet was Convolutional Neural Networks a special type of neural network that roughly imitates human vision. Over the years CNNs have become a very important part of many Computer Vision applications. So let's take a look at the workings of CNNs.

Background of CNNs

CNN's were first developed and used around the 1980s. The most that a CNN could do at that time was recognize handwritten digits. It was mostly used in the postal sectors to read zip codes, pin codes, etc. The important thing to remember about any deep learning model is that it requires a large amount of data

ADVANCED COMPUTER VISION DEEP LEARNING IMAGE ANALYSIS to train and also requires a lot of computing resources. This was a major drawback for CNNs at that period and hence CNNs were only limited to the postal sectors and it failed to enter the world of machine learning.

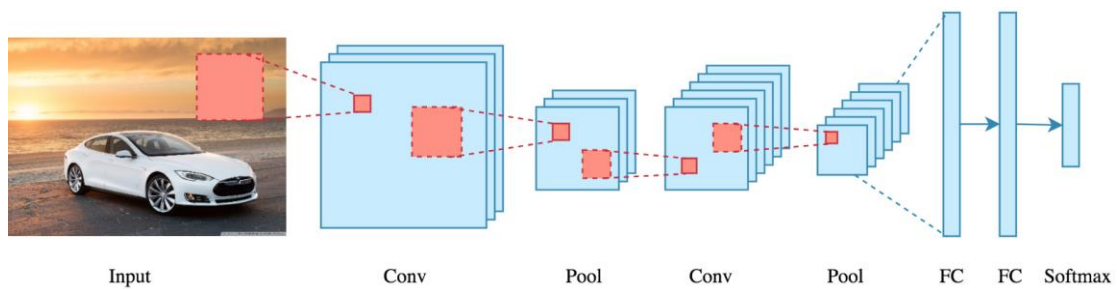


In 2012 Alex Krizhevsky realized that it was time to bring back the branch of deep learning that uses multi-layered neural networks. The availability of large sets of data, to be more specific ImageNet datasets with millions of labeled images and an abundance of computing resources enabled researchers to revive CNNs.

What exactly is a CNN?

In [deep learning](#), a **convolutional neural network (CNN/ConvNet)** is a class of [deep neural networks](#), most commonly applied to analyze visual imagery. Now when we think of a neural network we think about matrix multiplications but that is not the case with ConvNet. It uses a special technique called Convolution. Now in mathematics **convolution** is a mathematical operation on two functions that produces a third function

that expresses how the shape of one is modified by the other.



But we don't really need to go behind the mathematics part to understand what a CNN is or how it works. *Bottom line is that the role of the ConvNet is to reduce the images into a form that is easier to process, without losing features that are critical for getting a good prediction.*

How does it work?

Before we go to the working of CNN's let's cover the basics such as what is an image and how is it represented. An RGB image is nothing but a matrix of pixel values having three planes whereas a grayscale image is the same but it has a single plane. Take a look at this image to understand more Convolutional neural networks are composed of multiple layers of artificial neurons. Artificial neurons, a rough imitation of their biological counterparts, are mathematical functions that calculate the weighted sum of multiple inputs and outputs an activation value. When you input an image in a ConvNet, each layer generates several activation functions that are passed on to the next layer. The first layer usually extracts basic features such as horizontal or diagonal edges. This output is passed on to the next layer which detects more complex features such as corners or combinational edges. As we move deeper into the network it can identify even more complex features such as objects, faces, etc.

Based on the activation map of the final convolution layer, the classification layer outputs a set of confidence scores (values between 0 and 1) that specify how likely the image is to belong to a "class." For instance, if you have a ConvNet that detects cats, dogs, and horses, the output of the final layer is the possibility that the input image contains any of those animals.

What's a pooling layer?

Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to **decrease the computational power required to process the data** by reducing the dimensions. There are two types of pooling average pooling and max pooling. I've only had experience with Max Pooling so far I haven't faced any difficulties.

So what we do in Max Pooling is we find the maximum value of a pixel from a portion of the image covered by the kernel. Max Pooling also performs as a **Noise Suppressant**. It discards the noisy activations altogether and also performs de-noising along with dimensionality reduction. On the other hand, **Average Pooling** returns the **average of all the values** from the portion of the image covered by the Kernel. Average Pooling simply performs dimensionality reduction as a noise suppressing mechanism. Hence, we can say that **Max Pooling performs a lot better than Average Pooling**

Limitations

Despite the power and resource complexity of CNNs, they provide in-depth results. At the root of it all, it is just recognizing patterns and details that are so minute and inconspicuous that it goes unnoticed to the human eye. But when it comes to **understanding** the contents of an image it fails. Let's take a look at this example. When we pass the below image to a CNN it detects a person in their mid- 30s and a child probably around 10 years. But when we look at the same image we start thinking of multiple different scenarios. Maybe it's a father and son day out, a picnic or maybe they are camping. Maybe it is a school ground and the child scored a goal and his dad is happy so he lifts him.

These limitations are more than evident when it comes to practical applications. For example, CNN's were widely used to moderate content on social media. But despite the vast resources of images and videos that they were trained on it still isn't able to completely block and remove inappropriate content. As it turns out it flagged **a 30,000-year statue with nudity** on Facebook. Several studies have shown that CNNs trained on ImageNet and other popular datasets fail to detect objects when they see them under different lighting conditions and from new angles. Does this mean that CNNs are useless? Despite the limits of convolutional neural networks, however, there's no denying that they have caused a revolution in artificial intelligence. Today, CNN's are used in many **computer vision applications** such as facial recognition, image search, and editing, augmented reality, and more. As advances in convolutional neural networks show, our achievements are remarkable and useful, but we are still very far from **replicating the key components of human intelligence**.

Chapter 4

Machine learning

What are the 7 steps of machine learning?

7 Steps of Machine Learning

- Step 1: Gathering Data. ...
- Step 2: Preparing that Data. ...
- Step 3: Choosing a Model. ...
- Step 4: Training. ...
- Step 5: Evaluation. ...
- Step 6: Hyper parameter Tuning. ...
- Step 7: Prediction.

Introduction:

In this blog, we will discuss the workflow of a Machine learning project this includes all the steps required to build the proper machine learning project from scratch.

We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modelling steps that can help to improve the model performance.

Python Libraries that would be need to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit Learn
4. Matplotlib

Understanding the machine learning workflow

We can define the machine learning workflow in 3 stages.

1. Gathering data
2. Data pre-processing
3. Researching the model that will be best for the type of data
4. Training and testing the model
5. Evaluation

Okay but first let's start from the basics

What is the machine learning Model?

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if you give garbage to the model, you will get garbage in return, i.e. the trained model will provide false or wrong prediction

1. Gathering Data

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

We can also use some free data sets which are present on the internet. Kaggle and UCI Machine learning Repository are the repositories that are used the most for making Machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms, they also host competitions in which people can participate and get to test their knowledge of machine learning.

2. Data pre-processing

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.

What is data pre-processing?

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis.

Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

Why do we need it?

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects. Most of the real-world data is messy, some of these types of data are:

1. **Missing data:** Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).

2. **Noisy data:** This type of data is also called outliers, this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.

3. **Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of Data

1. Numeric e.g. income, age
2. Categorical e.g. gender, nationality
3. Ordinal e.g. low/medium/high

How can data pre-processing be performed?

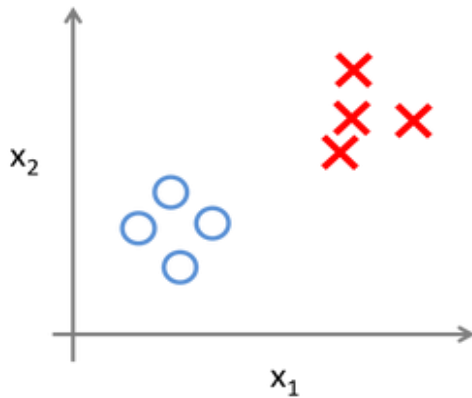
These are some of the basic pre — processing techniques that can be used to convert raw data.

1. **Conversion of data:** As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.
2. **Ignoring the missing values:** Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in the dataset.
3. **Filling the missing values:** Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.
4. **Machine learning:** If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.
5. **Outliers detection:** There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

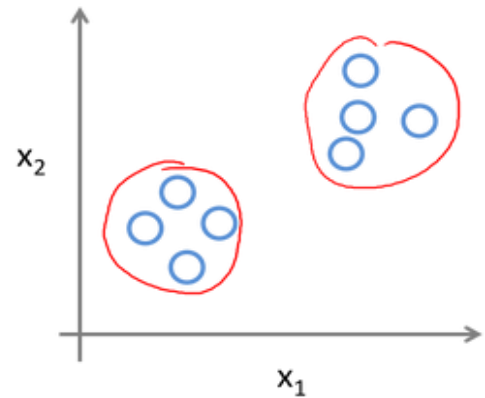
Researching the model that will be best for the type of data

Our main goal is to train the best performing model possible, using the pre-processed data.

Supervised Learning



Unsupervised Learning



Supervised Learning:

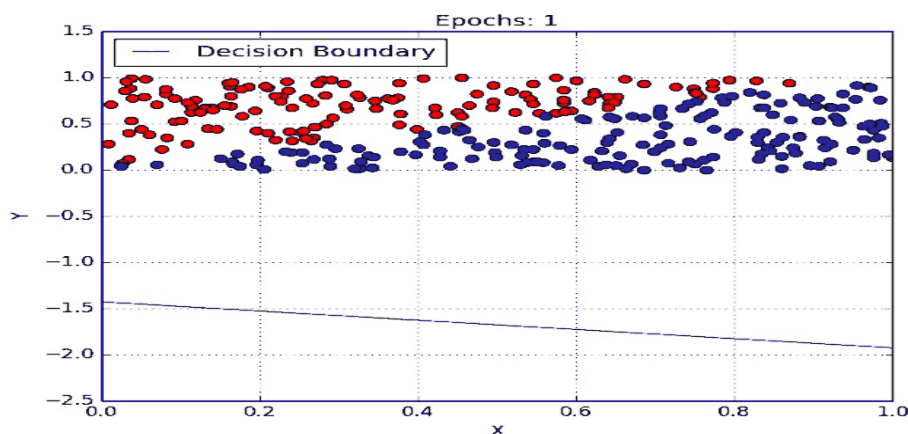
In Supervised learning, an AI system is presented with data which is labelled, which means that each data is tagged with the correct label.

The supervised learning is categorized into 2 other categories which are “**Classification**” and “**Regression**”.

Classification:

Classification problem is when the target variable is **categorical** (i.e. the output could be classified into classes — it belongs to either Class A or B or something else).

A classification problem is when the output variable is a category, such as “red” or “blue”, “disease” or “no disease” or “spam” or “not spam”.



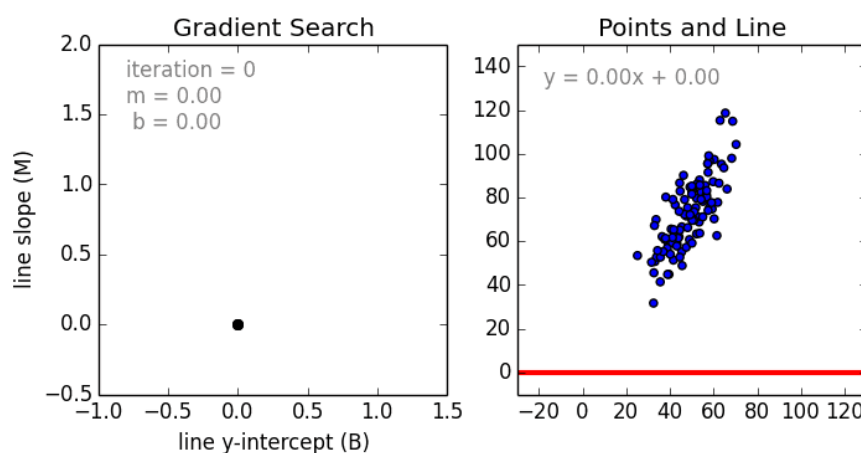
As shown in the above representation, we have 2 classes which are plotted on the graph i.e. red and blue which can be represented as 'setosa flower' and 'versicolor flower', we can image the X-axis as there 'Sepal Width' and the Y-axis as the 'Sepal Length', so we try to create the best fit line that separates both classes of flowers.

These some most used classification algorithms.

- **K-Nearest Neighbor**
- **Naive Bayes**
- **Decision Trees/Random Forest**
- **Support Vector Machine**
- **Logistic Regression**

Regression:

While a **Regression** problem is when the target variable is **continuous** (i.e. the output is numeric).



As shown in the above representation, we can imagine that the graph's X-axis is the 'Test scores' and the Y-axis represents 'IQ'. So we try to create the best fit line in the given graph so that we can use that line to predict any approximate IQ that isn't present in the given data.

These some most used regression algorithms.

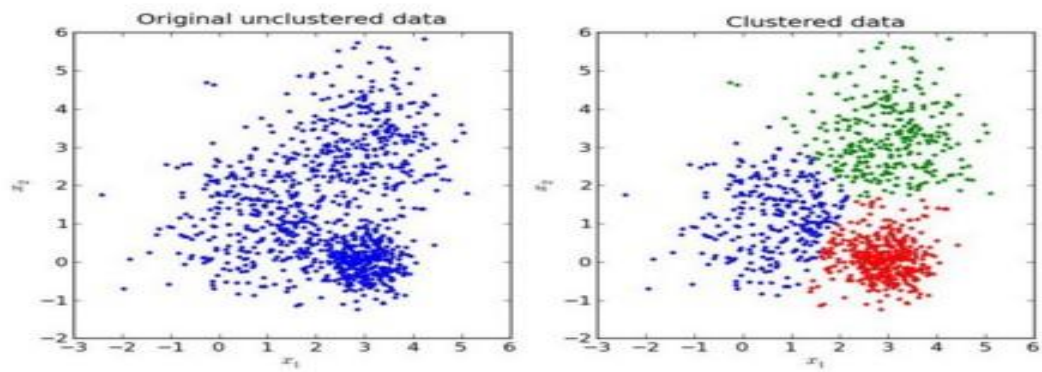
- **Linear Regression**
- **Support Vector Regression**
- **Decision Tress/Random Forest**
- **Gaussian Progresses Regression**
- **Ensemble Methods**

Unsupervised Learning:

The unsupervised learning is categorized into 2 other categories which are “**Clustering**” and “**Association**”.

Clustering:

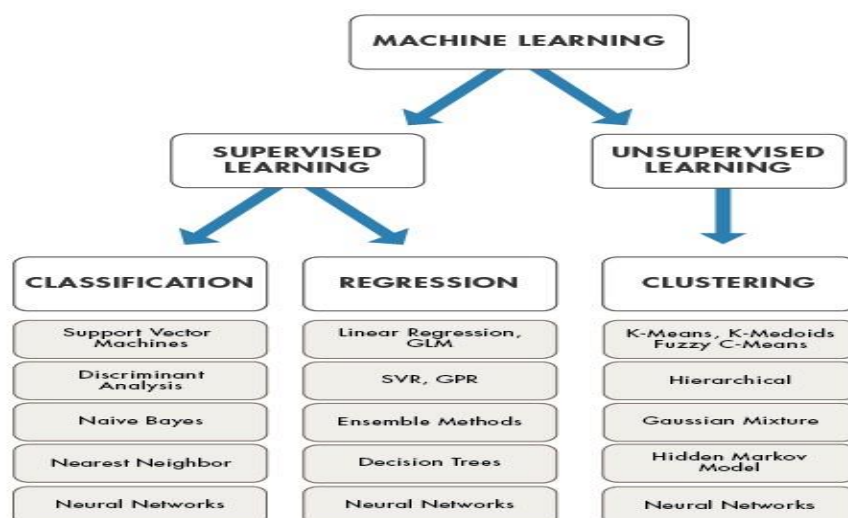
A set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.



Methods used for clustering are:

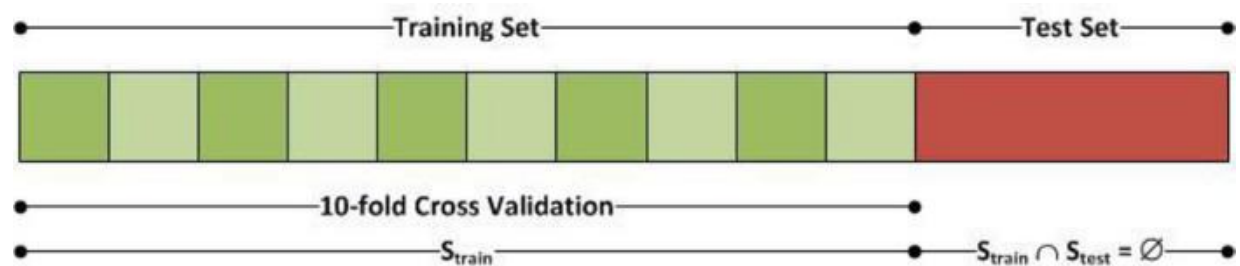
- Gaussian mixtures
- K-Means Clustering
- Boosting
- Hierarchical Clustering
- K-Means Clustering
- Spectral Clustering

Overview of models under categories:



4. Training and testing the model on data

For training a model we initially split the model into 3 three sections which are ‘**Training data**’ , ‘**Validation data**’ and ‘**Testing data**’. You train the classifier using ‘**training data set**’, tune the parameters using ‘**validation set**’ and then test the performance of your classifier on unseen ‘**test data set**’. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.

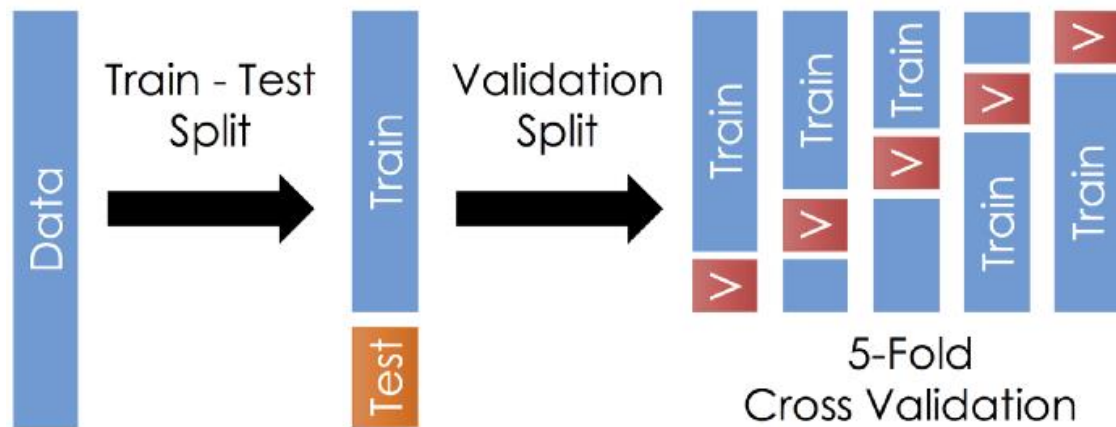


Training set:

The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning, that is to fit the parameters of the classifier.

Validation set:

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.



Once the data is divided into the 3 given segments we can start the training process.

In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use 'test set' instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

The model uses any one of the models that we had chosen in step 3/ point 3. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained. A confusion matrix has 4 parameters, which are '**True positives**', '**True Negatives**', '**False Positives**' and '**False Negative**'. We prefer that we get more values in the True negatives and true positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

True positives : These are cases in which we predicted TRUE and our predicted output is correct.

True negatives : We predicted FALSE and our predicted output is correct.

False positives : We predicted TRUE, but the actual predicted output is FALSE.

False negatives : We predicted FALSE, but the actual predicted output is TRUE.

We can also find out the accuracy of the model using the confusion matrix.

Accuracy = (True Positives + True Negatives) / (Total number of classes)

i.e. for the above example:

Accuracy = $(100 + 50) / 165 = 0.9090$ (90.9% accuracy)

5. Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives.

Conclusion

In this blog, we have discussed the workflow a Machine learning project and gives us a basic idea of how a should the problem be tackled.

Chapter 6

Conclusions:

- We have presented an approach and developed an algorithm for detecting malaria, automated malaria detection and quantification of malaria infection.
- Also, we developed a strategy to train with machine learning, adaptable to detection of malaria with other types of parasite and also discuss to increase the predictive value with results.

REFERENCES:

- [1] WHO, "Fact sheet: World Malaria Report 2016," in *World Health Organization*, world Health Organization, 2016. [Online]. Available: <http://www.who.int/malaria/media/world-malaria-report-2016/en>. Accessed: Jan. 25, 2017
- [2] Bloland PB (2001) Drug resistance in Malaria, WHO/CDS/ CSR/DRS/ 2001.4. World Health Organization, Switzerland, 2001
- [3] D Ghatge, C. Jadhav, and N. U. Rani, "AUTOMATIC DETECTION OF MALARIA PARASITE FROM BLOOD IMAGES," [Online]. Available: <http://ijact.org/volume4issue1/IJ0410050.pdf>. Accessed: Jan. 25, 2017
- [4] Ahirwar A., Pattnaik S., Acharya B., Advanced Image Analysis Based System for Automatic Detection and Classification of Malarial Parasite in Blood Images. *International Journal of Information Technology and Knowledge Management*. Jan-June 2012, Volume 5, No 1, pp. 59-64
- [5] Amit Kumar, Prof. Arun Choudhary, Prof. P. U. Tembhare, Prof. C. R. Pote. Enhanced Identification of Malarial Infected Objects using Otsu Algorithm from Thin Smear Digital Images. *International Journal of Latest Research in Science and Technology* Vol.1, Issue 2: Page No. 159-163, July-August (2012).
- [6] Nattkemper, T. W., Schubert, W., Hermann, T., & Ritter, H. (2004). A Hybrid System for Cell Detection in Digital Micrographs. In B. Tilg (Ed.), *Biomedical Engineering, Proc.* (Vol. 417). Innsbruck, Austria: ACTA Press.
- [7] CDC, "Frequently asked questions (FAQs)," CDC, 2016. [Online]. Available: <https://www.cdc.gov/malaria/about/faqs.html>. Accessed: Jan. 25, 2017.
- [8] L. Rosadoa, J. M. C. da Costa, D. Elias, and J. S. Cardoso, "Automated detection of malaria parasites on thick blood smears via mobile devices," 2016
- [9] N. A. Khan, H. Pervaz, A. Latif, and A. Musharaff, "Unsupervised identification of malaria parasites using computer vision," 2017.
- [10] D. Taylor, N. Green and G. Stout, 3rd ed., "Biological Science" The Press Syndicate of the University of Cambridge, 2002.
- [11] "malaria symptoms sore throat | Best Health Line", *Best Health Line*, 2017. [Online] Available: <http://besthealthline.com/malaria-symptoms-sore-throat/>. [Accessed: 29- Mar- 2017].
- [12] Gonzalez, R. and Woods, R. (2009). *Digital image processing*. 3rd ed. India: Dorling Kindersley India Pvt. Ltd.

- [13] Solomon, C. and Breckon, T. (2013). *Fundamentals of digital image processing*. Hoboken, N.J.: Wiley, [Sec. 10.10].
- [14] Colour Space Conversions. (1998). Adrian Ford (ajoec1@wmin.ac.uk <defunct>) and Alan Roberts (Alan.Roberts@rd.bbc.co.uk)., pp.5,10-11.
- [15] Mathworks.com. (2017). *Measure properties of image regions - MATLAB regionprops - MathWorks United Kingdom*. [online] Available at: <https://www.mathworks.com/help/images/ref/regionprops.html> [Accessed 27 Jun. 2017].
- [16] World Health Organization, Malaria Report 2013.