Artificial Intelligence Project

Malaria Detection

Submitted by

TEAM-12

Preeti Prasad-00501032017

Ekta Pal -02401032017

Vandana -03201032017

Lamha - 04101032017

Riya Raj Kanojia -06601032017

Kriti -07001032017

Under the Supervision of

Mr. Rishabh Kaushal

Assistant Professor

Department of Information Technology

Indira Gandhi Delhi Technical University for Women

Kashmere Gate, Delhi – 110006

STUDENT UNDERTAKING

This is to undertake the work titled in this Minor Project Report as part of 3rd Semester in B.Tech. (Information Technology) with specialization during March – May, 2020 under the guidance of Mr. Rishabh Kausal(Assistant Professor)of Artificial Intelligence This is our original work.

The report has been written by us in our own words and not copied from elsewhere. This report was submitted to plagiarism detection software on (25-05-2020) and percentage similarity found was 4-5% , similarity report attached as Appendix.

Anything that appears in this report which is not my original has been duly and appropriately referred / cited / acknowledged. Any academic misconduct and dishonesty found now or in future in regard to above or any other matter pertaining to this report shall be solely and entirely my responsibility. In such a situation, I understand that a strict disciplinary action can be undertaken against me by the concerned authorities of the University now or in future and I shall abide by it.

DEPARTMENT OF INFORMATION TECHNOLOGY INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN
KASHMERE GATE, DELHI - 110006

CERTIFICATE

This is to certify that the work titled submitted by in this project report as part of 3rd Semester in B.Tech. (Information Technology) with specialization during March – May, 2020 was done under my guidance and supervision.

This work is her original work to the best of my knowledge and has not been submitted anywhere else for the award of any credits / degree whatsoever. The work is satisfactory for the award of Minor Project credits.

ACKNOWLEDGEMENT

The success and final outcome of this project needed a lot of advice and assistance from many people and we are privileged to have this all along the completion of our project.Everything we have done so far has become possible due to this supervision.We are very grateful to each one of them.

We are extremely thankful to Mr.Rishabh Kaushal for providing us with this opportunity of doing an AI project in malaria detection guiding us throughout the project. We are extremely thankful to him for providing his support and guidance.

Contents

Chapter 1

Introduction

Malaria is an infectious disease caused by microorganisms and it poses a major threat to the global health zone. In the tropical and subtropical countries including India, malaria has been a challenge, which really needs a quick and precise diagnosis to put a stop or control the disease. The conventional microscopy method has some shortcomings which includes time consumption and reproducibility. Many of the alternative methods are expensive and it's not readily accessible to the developing countries like India that need them.

1.1 Problem Statement

Malaria has been a huge problem in India and its diagnosis using conventional methods is very costly and not efficient. Therefore, automation of the evaluation process in the diagnosis of malaria is highly important. In this project, we are describing a method that uses image analysis and classifies the image of blood

cells. This process involves two main phases. First, pre-processing, where the images are corrected for luminance and transformed to a constant color space. Second, feature extraction using a support vector machine (SVM) for the diagnosis of the disease in the red blood cell. In this method features and parameters are computed from the data obtained by the digital images of the blood cells and is given as input to SVM which classify the cell as the infected one or otherwise.

1.1.1 Objective

- Predicting whether an image sample of red blood cells is infected or not?
- Building an accurate and cost effective methodology to detect malaria in blood samples

Chapter 2

Literature Survey

2.1 Research questions

1.Will this project help in predicting the group of people whose lives are more endangered by Malaria based on their gender, age and living conditions?

2.Are we able to find the severity of malaria disease in a patient by the help of this project?

Chapter – 3

Proposed Methodology

3.1 Dataset

Number of parasitized - 151

Number of uninfected- 50

Number of Instances- 22046 images

Number of Attributes- 4

Label Information- Label Present

The Dataset are:

Test data

http://www.codeheroku.com/static/workshop/datasets/malaria_detection/test.csv

Train data:

http://www.codeheroku.com/static/workshop/datasets/malaria_detection/train.csv

3.2 Attribute Description

The method demands experience in analyzing the variability in SIZE ,BACKGROUND, ANGLE, and POSITION of the region of interest (ROI) on the pictures.

3.3 Details concerning the Organization

This page hosts a repository of segmental cells from the skinny blood smear slide pictures from the protozoal infection guard analysis activity to scale back the burden for microscopists in resource-constrained regions and improve diagnostic accuracy, researchers at the Lister Hill National Center for medical specialty Communications (LHNCBC), a part of National Library of medication (NLM), have developed a mobile application that runs on a typical robot smartphone hooked up to a traditional microscope. Giemsa-stained skinny blood smear slides from a hundred and fifty P. falciparum-infected and fifty healthy patients were collected and photographed at urban center Medical school Hospital, Bangladesh. The smartphone's integral camera non heritable pictures of slides for every field} of view.The pictures were manually associated by an skilled slide reader at the Mahidol-Oxford medical specialty analysis Unit in Krung Thep, Thailand. The de-identified pictures and annotations area unit archived at NLM (IRB#12972). we have a tendency to apply a level-set based mostly algorithmic program to discover and section the red blood cells.

3.4 Data Pre-processing

3.4.1 Knowledge Files Combining

The CSV files containing the Patient-ID to cell mappings for the parasitized and antiseptic categories. The CSV file for the parasitized category contains 151 patient-ID entries. The slide pictures for the parasitized patient-ID "C47P8thinOriginal" area unit scan from 2 completely different magnifier models (Olympus and Motif). The CSV file for the antiseptic category contains 201 entries since the conventional cells from the infected patients' slides additionally build it to the conventional cell class (151+50 = 201).

These files were combined to create one dataset.

3.4.2 knowledge cleanup

1. scan original image

2. Convert to grayscale

3. Contour Detection

4. Get areas of largest contour

3.5 Feature Computation

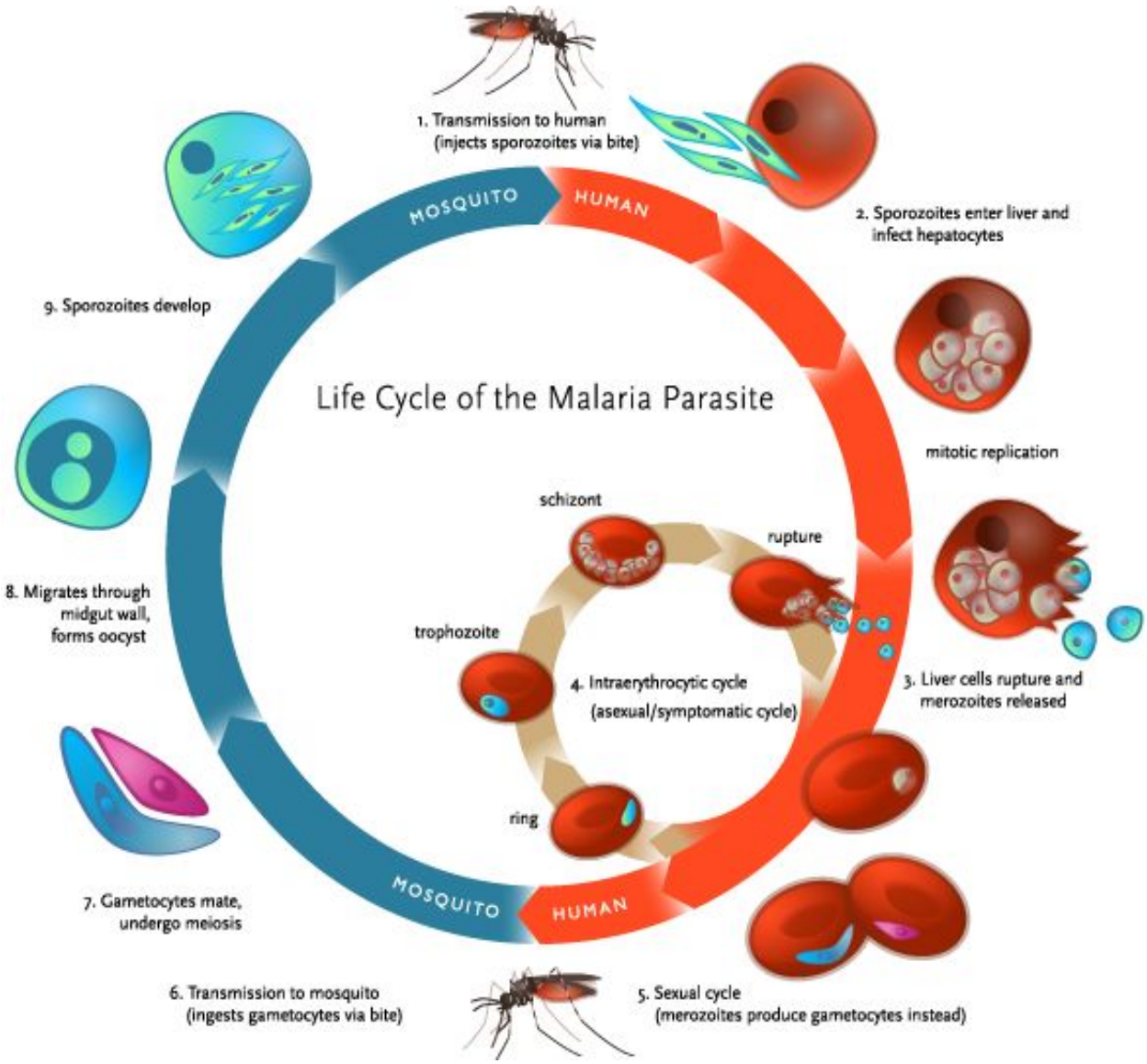
Direct Features:

These features are already present in the dataset as attributes, so no computation is required as such.

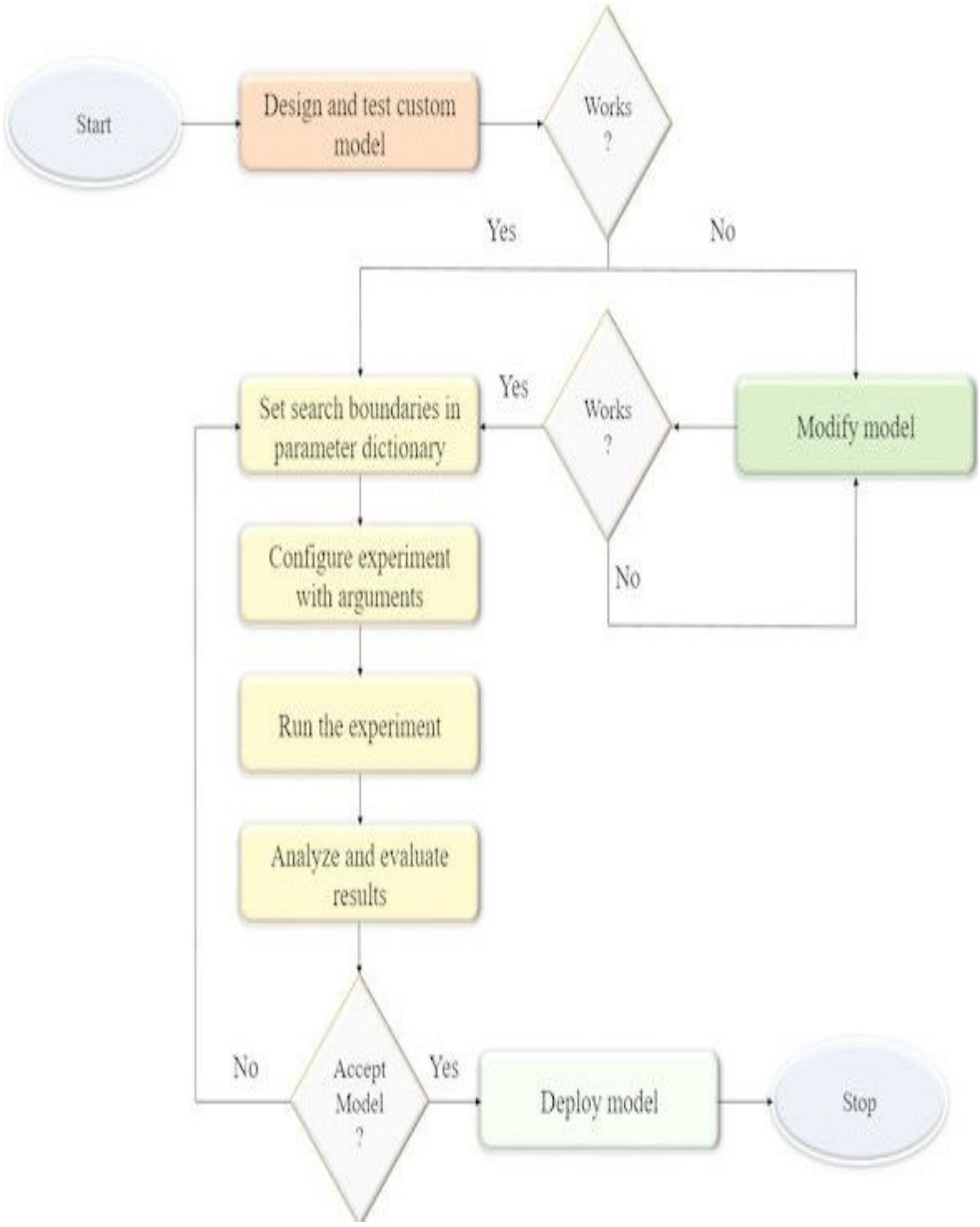
Chapter – 4

Diagrams and Graph

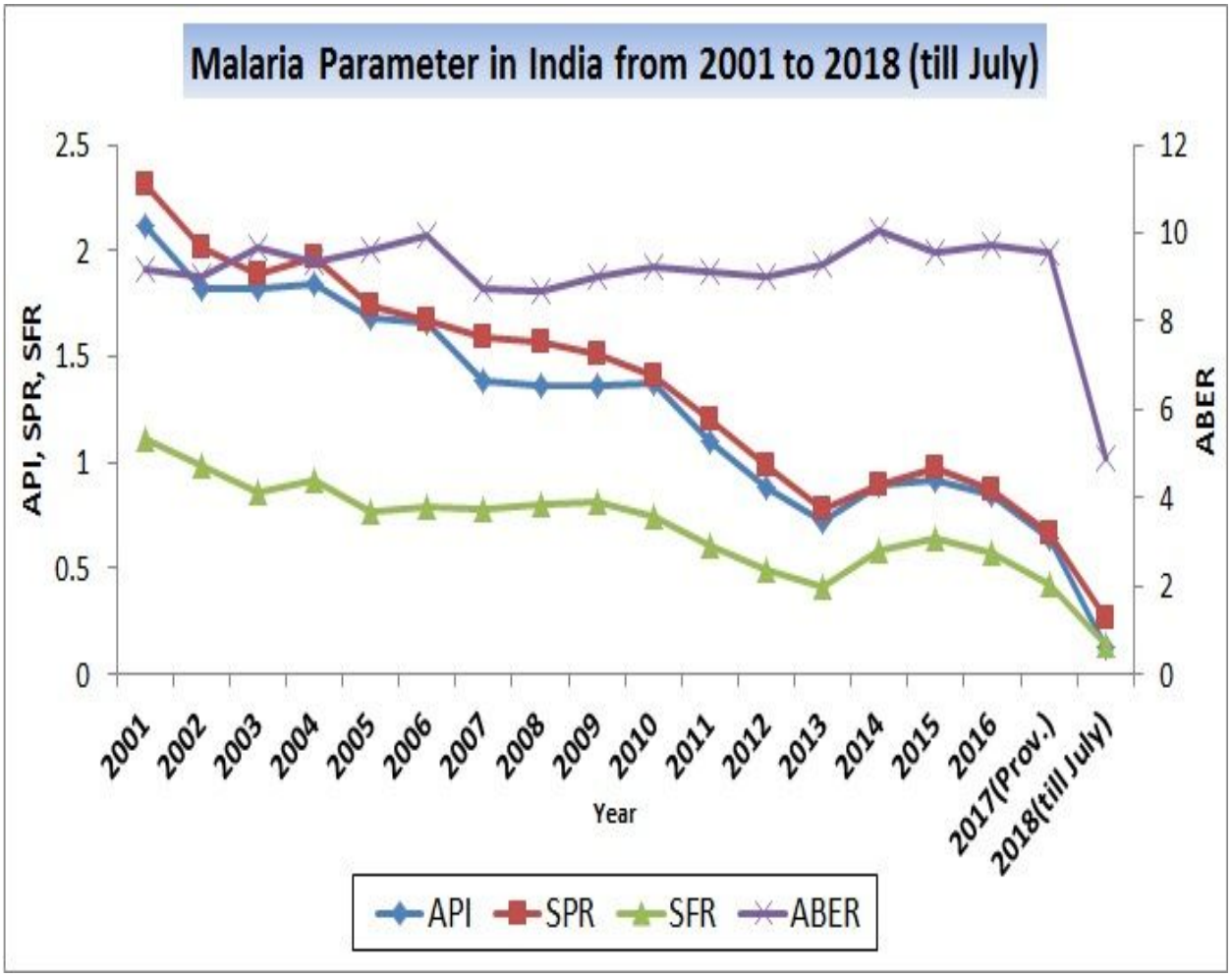
4.1 LIFE CYCLE OF MALARIA PARASITE



4.2 Process flow diagram for optimizing the hyper parameters of the custom CNN model



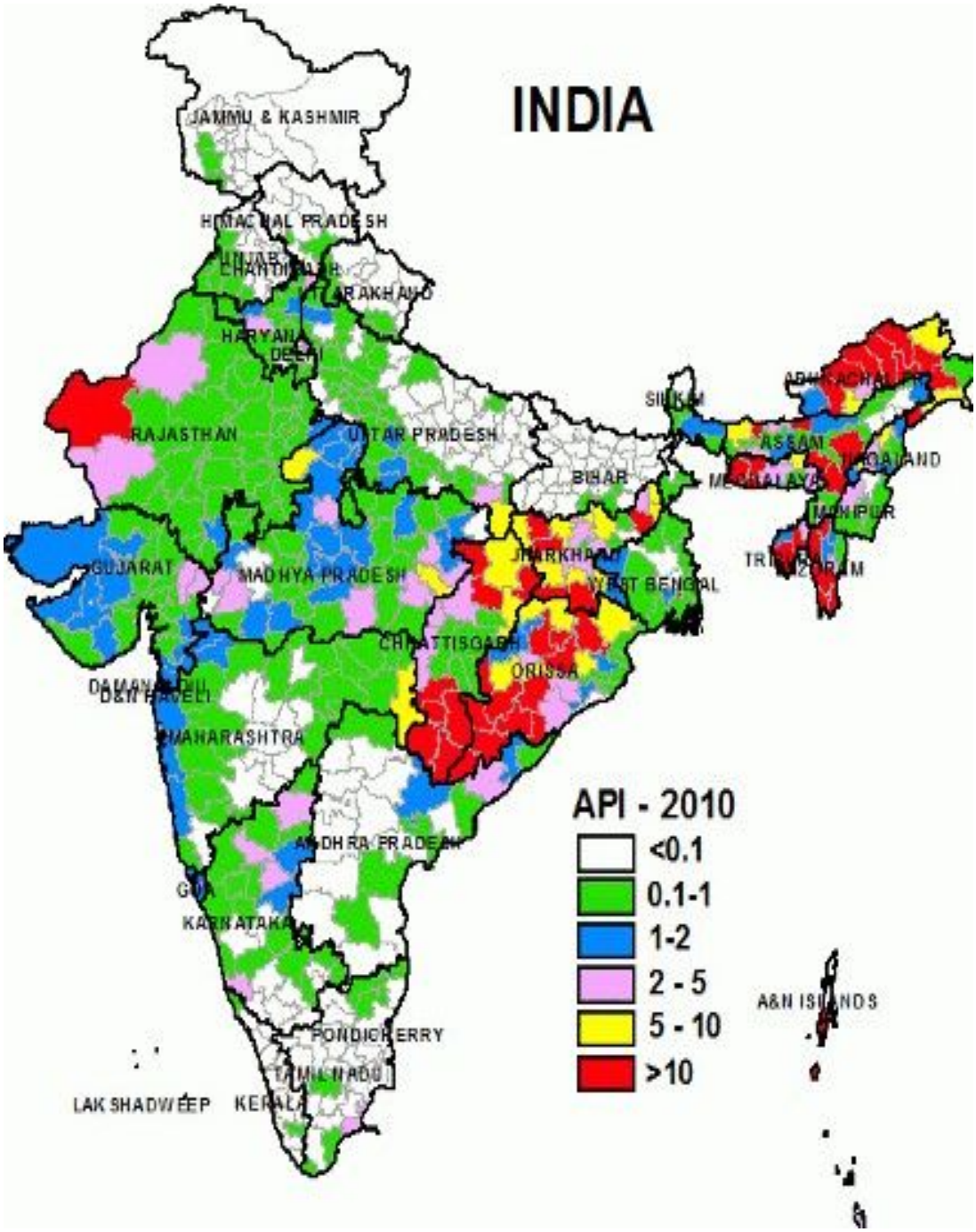
4.3 Trends of malaria incidence in India from 2001 to 2018



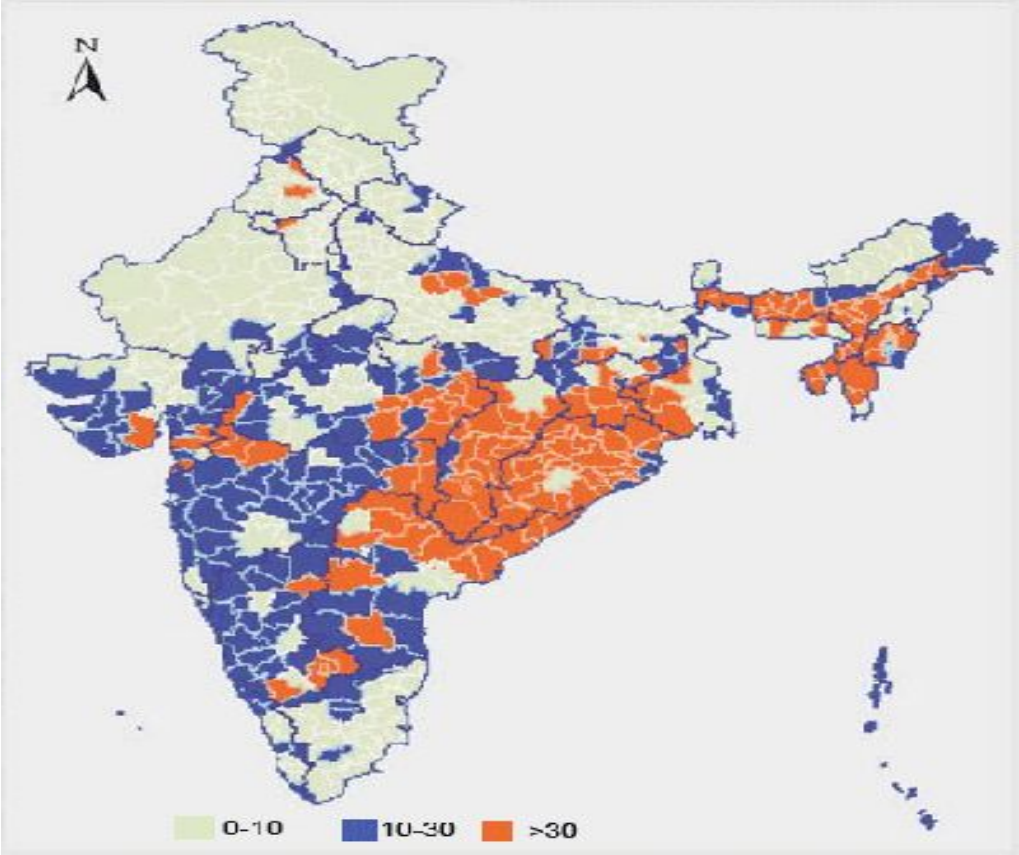
Where, ABER=annual blood examination rate

API= annual parasite incidence

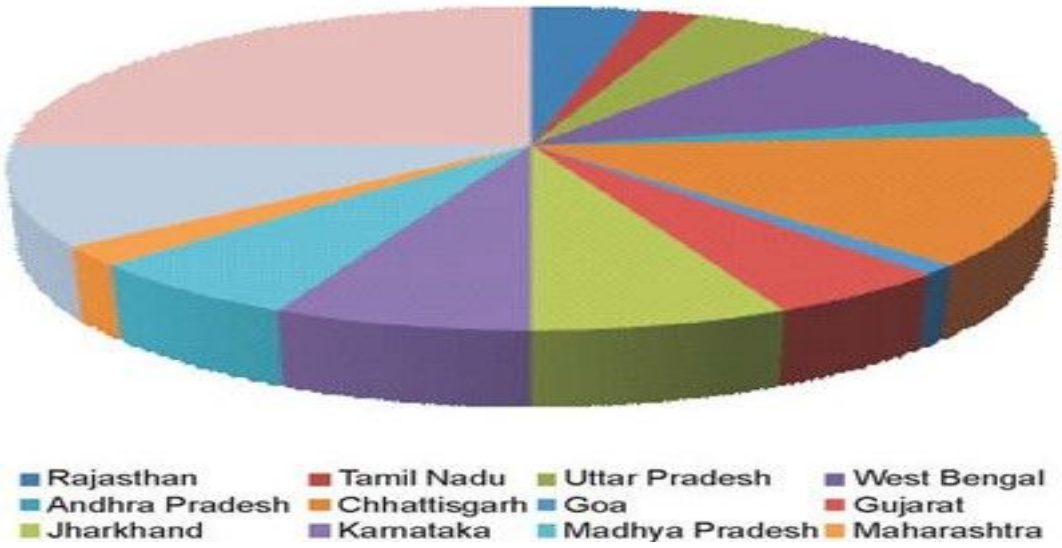
4.4 Distribution of Malaria Incidence in India according to API


INDIA
JAMMU & KASHMIR
HIMACHAL PRADESH
HARYANA
DELHI
RAJASTHAN
UTTAR PRADESH
BIHAR
SIKKIM
ASSAM
MEGHALAYA
MANIPUR
GUJARAT
MADHYA PRADESH
JHARKHAND
WEST BENGAL
CHHATTISGARH
ORISSA
MAHARASHTRA
ANDHRA PRADESH
GOA
KARNATAKA
PONDICHERRY
TAMIL NADU
KERALA
LAKSHADWEEP
A&N ISLANDS
API - 2010
<0.1
0.1-1
1-2
2 - 5
5 - 10
>10

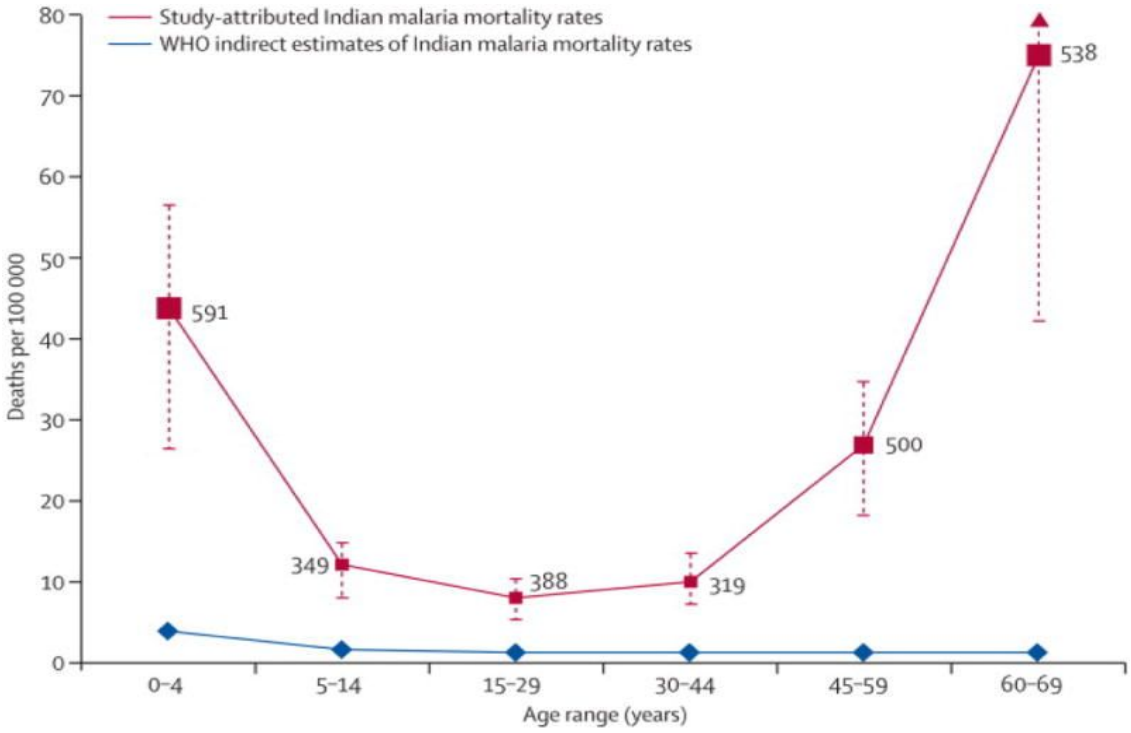
4.5 Proportion of P. falciparum distribution in India



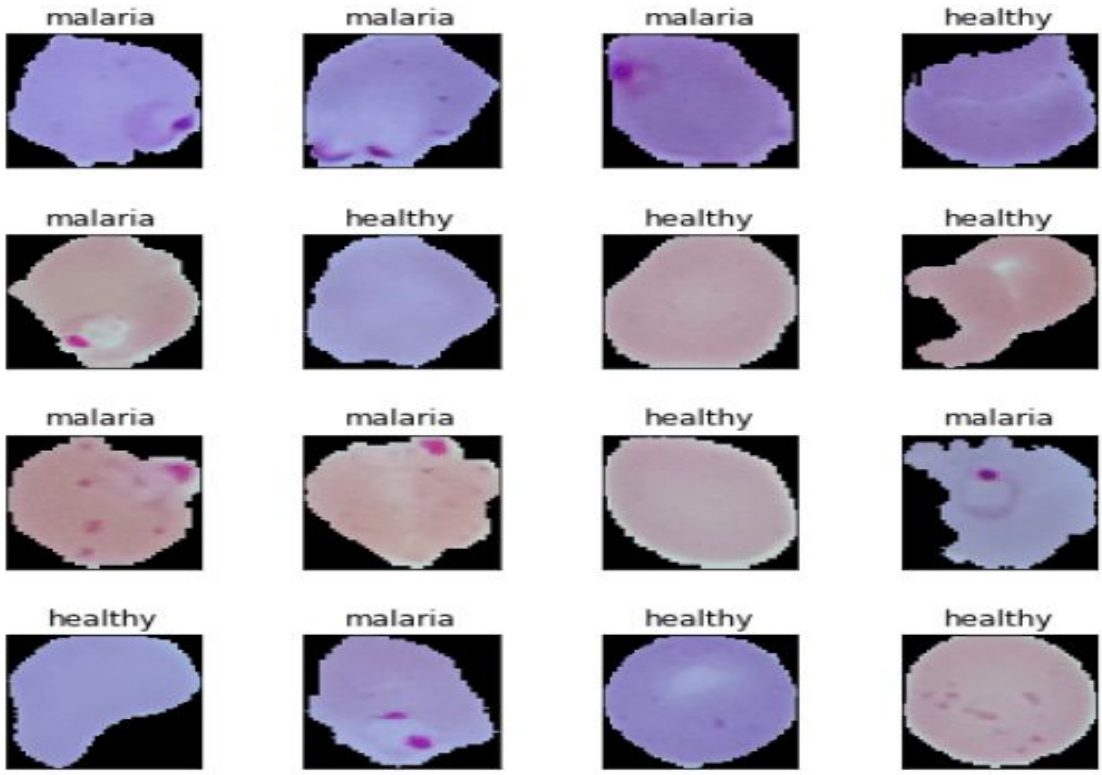
4.6 Contribution of different states to malaria in India



4.7 Age and sex distribution of malaria mortality in India



4.8 Differences between malaria and healthy cell images



Chapter – 5

Algorithms

5.1 Support Vector Machine

5.1.1 What is a Support Vector Machine?

A support vector machine is one of the classical algorithms which comes under supervised learning .This is mostly used for classification problems including two groups. It can also be used for Regression.

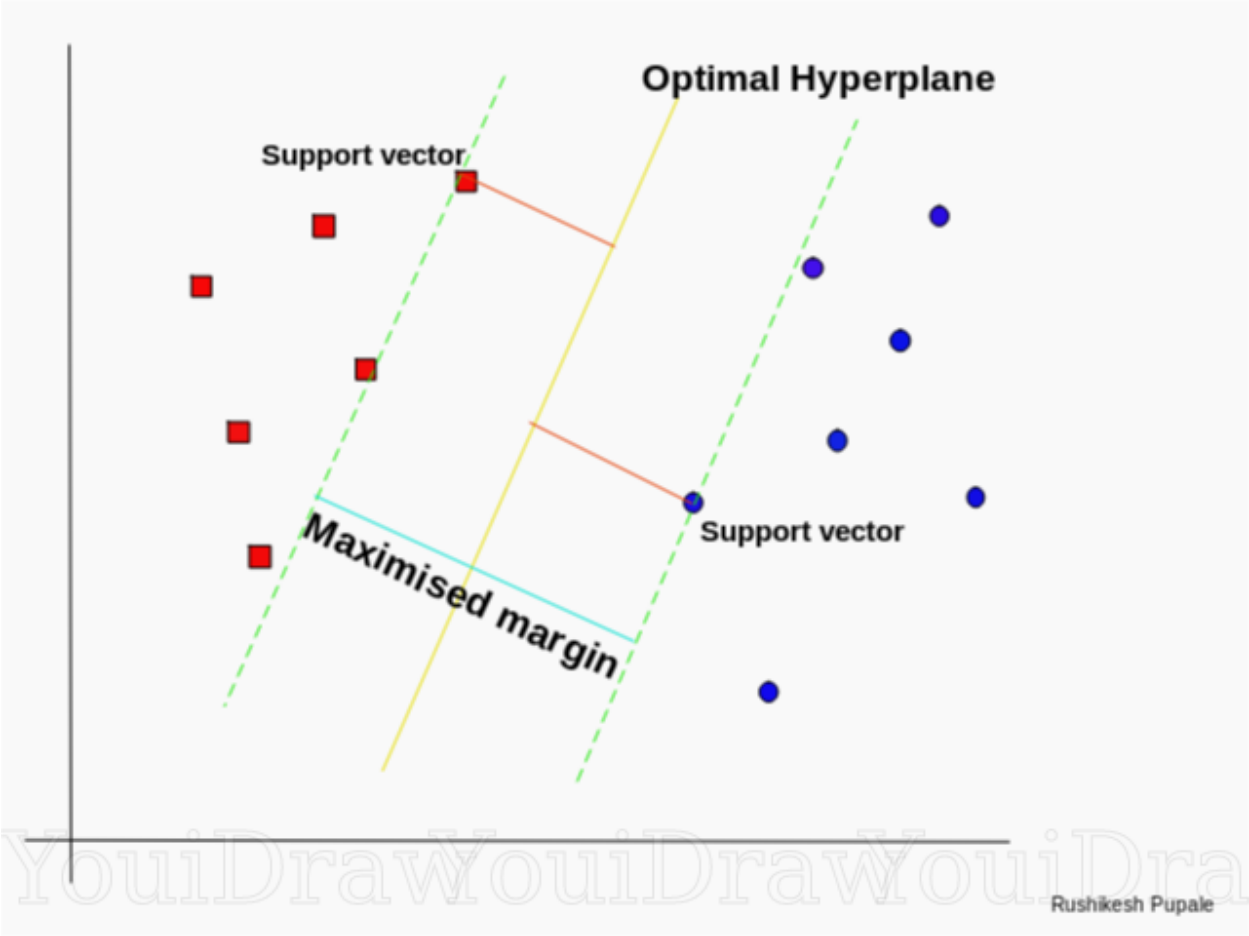
It is a better approach as it is fast and dependable .It works very precisely with a limited set of data .It gives optimal solutions .So,we are using this approach as we have a limited data set(fixed) and it is a more preferred approach over Naive Bayes.

Terms used in SVM

Hyperplane-line which separates infected and uninfected image sets(in our project)

Margin-It is maximum distance between two support vectors.

Support vector-closely present images on the left and right side of hyper plane



5.1.2 General Idea of how does a SVM works-

SVM is a linear model and can be used for classification as well as regression but here we are mainly focusing on classification that is we are solving a classification problem using SVM as we have a dataset which is divided into two groups one with the infected RBCs and one with the uninfected RBCs.It's working is simple as it creates a hyperplane or a line which separates the data of the two groups.Separator or the line which separates the two groups is termed as a hyperplane the data is transformed in a way so that we can get the hyperplane it includes a number of calculation and detail derivations we will not go into that.

Here one question arises .How can we decide which is the most suitable hyperplane ? or what are the criterias of choosing a hyperplane?

We will choose a hyperplane such that it maximizes the margins from both groups that is the distance from the hyperplane and is the largest from both the groups.

Let us understand svm with the help of an example. We have data in the form of squares and circles. This data is labelled.
Now the question arises why is it labelled?
So,the ans is because as we have mentioned it earlier svm is a supervised learning
Now we will send this labelled data to the training model to build our model
After preparing our model we will send it for testing for prediction of new upcoming data to find its class of belonging that is providing us our output and hence we can find out data which is in training phase belongs to quadrilateral class or to the circle class

Approach of SVM used in our project

In our project we have performance contour detection which is a feature extraction technique which will essentially mark all the bounded regions

Then we will get areas of 5 largest contours .
For example-for an infected image we will get 4 contours or 3 smaller patterns but for an uninfected image we will get only 1 bigger contour which would be itself the cell boundary.

5.3 CNN (Convolutional Neural Networks) ALGORITHM

For project malaria detection for feature extraction and classification technique we could also use CNN which would be used for both supervised and unsupervised but here in our project we performed CNN using supervised learning

What is CNN ?

CNN is a type of feed forward artificial neural network in which connectivity pattern between its neurone is prompt by organisation of the one word, here visual cortex
the visual cortex has small region of cell that are sensitive of specific region of the visual field, some individual neural cell in the brain respond only when a job of a certain orientation happen
Let us take an example of some neuron fires when they are exposed to vertical edges and some when shown horizontal or diagonal edges.

FULLY CONNECTED LAYER:
This is the final layer ,where we will perform the actual classification.Here we will take filtered and shrinked images and then put them into a single list.

5.3.1How CNN works

Basically three layer:

- ❖ Convolution layer
- ❖ Relu layer
- ❖ Pooling
- ❖ Fully connected

CNN compare the using piece by piece In pieces it looks for features
By rough feature matches in roughly the same position into image
CNN gets are not better at Singh similarity then hole image matches
In this layer we need to revcore every negative value from the filtered image and replace it with the zero.

CONVOLUTIONAL LAYER:

A convolutional neural network is a type of deep learning and it is most commonly used in analyzing visual imagery.we can also call it shift invariant or space invariant artificial neural networks.
The convolutional layer is the core building block of a CNN .Here the layers parameter consists of a set of learnable filters also called as kernels.Kernel have a small receptive field,but could be extended via full depth of the input volume.
As we go deeper in the convolutional neural network containing convolutional layers,the filters do dot product of the input of the previous convolutional layer .So,basically they are taking a smaller coloured edges and then making larger pieces from them.The input exists in the window of pixels with the channels having depth. This is the same with the output which is considered as a 1 by 1 pixels "window:".

ReLU (Rectified Linear Unit) :

It transforms our function and only activates a node if the input is above a certain quantity ,when the input is below zero ,pour output is zero ,but when our input rises above a certain threshold value ,linear relationship is shown by it with the dependent variable.

POOLING LAYER:

In this layer we shrink the image stack into a smaller size.
Steps of doing this:
1.pick a window size ,usually 2 or 3
2.pick a stride(usually 2)
3.Walk your window across your filtered image .
4.For each and every window use the maximum values.

FULLY CONNECTED LAYER:
This is the final layer ,where we will perform the actual classification.Here we will take filtered and shrinked images and then put them into a single list.

5.3.2UNDERSTANDING CNN in our Project MALARIA DETECTION

Infected and Uninfected Image Identifier :
infected images(data set 1)
uninfected images(dataset 2)

INPUT--->CONVOLUTIONAL NEURAL NETWORKS ----->DETECTING INFECTED OR UNINFECTED IMAGE

1. Implementing the use -Case:
2. Download the dataset
3. Function to encode the labels
4. Resize the image 50*50 pixels and read it as a grayscale image.
5. Split the data for training and testing phase
6. Reshape the data appropriates for Tensorflow
7. Build the approach model
8. Now calculate loss function,it is categorical cross entropy use

9. Adam as optimizer with learning rateset
10. Train the Deep Neural Network for 10 epach.
11. Make prediction of result

5.3.3 Convolutional neural network have higher accuracy,precision and recall than SVm

1.CNN build
their own features from raw signal.It opposed to other algorithm which uses vector representations and where every component usually makes sense own its own.Here in cnn pixels don't have meaning outside the context ,but together they can contain more information about the object on a image than a bunch of its properties that you feed into support vector machine.
2.CNN uses strong priors which is the primary property of max pooling layers.Here properties like good generalization and invariance to local fluctuations makes them super scalable.
3.CNN has spatial feautres.It can be its strength if the context of the feature is local and it can be their weakness if our context is distributed.A sum of one-hot encodings is more difficult in handling with CNNs networks but its very easy in case of decision tree.
4.CNN stores much more information in the form of parameters than the other methods .Trees in RFs can tell you about features importance if you are dealing with interpretable features .but convents have a better approach in utilization of enormous amounts of parameters.
ADVANTAGES OF CNN OVER SVM:
Neither is inherently better approach ,but they each have their own advantages and disadvantages .CNN is a good candidate for image recognition.We could use CNN for sure in sequencing of data ,but they shine in going through huge amount of image and finding no-linear correlation.whereas here SVM face problems in predicting class labels when the size of class labels is huge.Also its quite hard task to parallelize SVM but the CNN technique or architecture support it.

Chapter – 6

Conclusion

This report is on malaria detection. We have read pre - work on malaria detection. After reading and understanding we designed some general questions and problems we are facing in India because of malaria and how our machine learning algorithm is going to provide an easy solution for malaria detection. We are going to use non - linear processing SVM using a kernel algorithm and CNN for classification of infected and non-infected cells. We studied about support vector machines, studies on malaria and its growth in India and tried to analyse it with the help of graphs and charts. After analyzing , we figured out instances , attributes and labels.Then, we have done the preprocessing of the images so that we can easily detect contour.After contour detection, we are classifying infected cell images from uninfected cell images by determine number of contours, shape of and size of contours.

We are increasing the accuracy score eventually up to 20epoch Ya could not be exactly the same because of the initial weight out sarcastic gradients descent algorithm we are never guaranteed that we reach the global Optima initial weight if they are random we start with a random. It is pretty much random with 12 epoch over where we are getting about 96 %accuracy which is ready good but this is our training accuracy this is not same as test accuracy we need to make short that we evaluate our model on a test data

Precaution

This will evaluate a model test data i.e test accuracy is 95.5 %which is not so bad of any classification tasks

When we get actual [1,0] which is first class i.e parasite class (1 predict the parasite image)probability of class one is [9.99 2.11] 9.99 can predict class 1 and 2.11 which is much smaller
When we get the actual [0,1] which is uninfected class (0 predict the uninfected image) predicted
This is not infected and higher and seems like we are doing pretty good for classification test

References

1.https://www.malariasite.com/malaria-india/
2.https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria
3.https://en.m.wikipedia.org