iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4437

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Classifying Malaria Cell Images Dataset using Machine Learning Algorithms

Dr. S. Jessica Saritha[1], Puvvula Spandana[2], Manukonda Alfred Raju[3], Anupatti Ediga Jagadeesh Goud[4]
[1]*Assistant Professor*, [2, 3, 4]*Student, Dept of CSE, JNTUACEP, Pulivendula, AP, India,*

Abstract: ***In today's world, detecting malaria has become a very common phenomenon. The project we have selected is taken from the area of medical sector and Health organizations. A large number of people are affected by malaria. But all these people are not correctly predicted whether they are affected by malaria or not. Every year, we read about a number of cases where people died with malaria. The risk associated with making decision on predicting a person affected with malaria accurately is immense. So, the idea of this project is to gather data from multiple data sources and use machine learning algorithms on this data to extract important information and predict whether a person is affected with malaria or not.***
Keywords: Machine learning algorithm, Malaria cell image predictions, Standard CNN, ResNet50.

I. INTRODUCTION

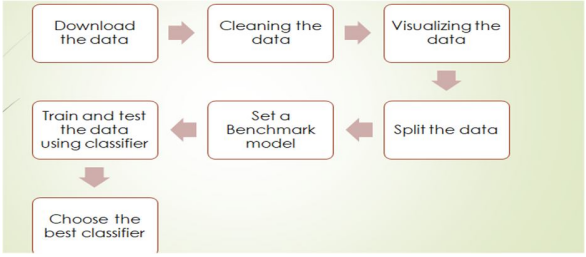
Malaria, sometimes called the "King of Diseases", is caused by protozoan parasites of the genus *Plasmodium*. The most serious and sometimes fatal type of malaria is caused by *Plasmodium falciparum*. The other human malaria species, *P. vivax, P. ovale, P. malariae*, and sometimes *P. knowlesi* can cause acute, severe illness but mortality rates are low. Malaria is the most important infectious disease in tropical and subtropical regions, and continues to be a major global health problem, with over 40% of the world's population exposed to varying degrees of malaria risk in some 100 countries. It is estimated that over 500 million people suffer from malaria infections annually, resulting in about 1-2 million deaths, of whom 90% are children in sub- Saharan Africa. The number of malaria cases worldwide seems to be increasing, due to increasing transmission risk in areas where malaria control has declined, the increasing prevalence of drug resistant strains of parasites, and in a relatively few cases, massive increases in international travel and migration. The need for effective and practical diagnostics for global malaria control is increasing, since effective diagnosis reduces both complications and mortality from malaria. Differentiation of clinical diagnoses from other tropical infections, based on patients' signs and symptoms or physicians' findings, may be difficult. Therefore, confirmatory diagnoses using laboratory technologies are urgently needed. Two machine learning algorithms are discussed on the currently available diagnostic methods for malaria in many settings, and assesses their feasibility in resource-rich and resource-poor settings. They are Standard CNN, ResNet50. The machine learning algorithm Standard CNN performance is used to compare with other algorithm ResNet50 performance through accuracy.

A. Motivation

We will use this project in the detection of malaria using cell images i.e., either the image is parasitized or uninfected. We have developed this project by using image classification algorithms that are involved in machine learning. Once we give the image of the cell, we will be able to predict whether the cell image is infected by malaria or not.

B. Proposed Work

By using data mining techniques predicting the existence of malaria is a time-consuming task. So, in the proposed system we will use different image classification models to find the accuracy given by each model and finally selects the best model which gives the highest accuracy. Some of the image classification models which used are Sequential(Standard CNN), ResNet50. Task flow involved during developing of this application is



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

II. DATA ANALYSIS

The objective of data analysis step is to increase the understanding of the problem from the data. There are two approaches to describe a given dataset. Summarizing and Visualizing data.

A. Data Exploration

Data Pre-processing:

The pre-processing done to "Prepare data" for the classification task, the notebook consists of following steps:

1) The list of images is randomized to get good redistribution of data when applying split.
2) The images are divided into a training set and a validation set (10% of original train set) to avoid overfitting
3) The images are resized into 3D tensor with shape of width = 50, height = 50, channels = 3 shaped arrays, so that every image will have same dimensions.
4) All images are then normalized by dividing with 255. 6.
5) All labels are converted to 2 categories by using keras.utils.to_categorical: in order use for categorical classification. Cell images are not reduced to a single size, they have different proportions.

B. Data Set

https://www.kaggle.com/iarunava/cell- images-for-detecting-malaria/home
This dataset contains 27,558 cell images.

C. Categories

As it is a categorical data, we have two labels. They are Parasitized and uninfected.

III. ALGORITHMS AND TECHNIQUES

A. Artificial Neural Networks (ANNs)

ANNs are computing systems vaguely inspired by the biological neural networks that constitute animal brains and humans. these systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. A typical brain contains something like 100 billion miniscule cells called neurons (no-one knows exactly how many there are and estimates go from about 50 billion to as many as 500 billion).

Each neuron is made up of a cell body (the central mass of the cell) with a number of connections coming off it: numerous dendrites (the cell's inputs—carrying information toward the cell body) and a single axon (the cell's output—carrying information away). Neurons are so tiny that you could pack about 100 of their cell bodies into a single millimetre. Inside a computer, the equivalent to a brain cell is a tiny switching device called a transistor. The latest, cutting-edge microprocessors (single-chip computers) contain over 2 billion transistors; even a basic microprocessor has about 50 million transistors, all packed onto an integrated circuit just 25mm square.

The Set of connected neurons organized in layers:

1) *Input layer*: brings the initial data into the system for further processing by subsequent layers of artificial neurons. Like dendrites in human neuron.
2) *Hidden layer*: a layer in between input layers and output layers, where artificial neurons take in a set of weighted inputs and produce an output through an activation function. Like nucleus in human neuron.
3) *Output layer*: the last layer of neurons that produces given outputs for the program. Like axon in human neuron.

NOTE: No one how they process information inside. A typical Artificial neural network consists of 3 parts:

B. ResNet50

ResNet-50 is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 50 layers deep and can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224.

C. Real Time Example

Object recognition,Visual recognition tasks, Face recognition etc.

Advantages:

1) Reduce misclassifications which saves Human life.
2) Prediction is done with more accuracy.
3) It minimizes time.
4) Most suitable for high recall problems.
5) All items in the data set are trained without missing.

D. Parameters

Keras.applications.resnet.ResNet50(include_top=**True**,weights='imagenet',input_tensor=**None**,input_shape=**None**,pooling=**None**, classes=1000)

E. Standard CNN Model

Standard Convolutional Neural Networks (CNNs) are neural networks with architectural constraints to reduce computational complexity and ensure translational invariance. Standard CNNs have two non-linearities:

Pooling layers and ReLU functions.Pooling layers consider a block of input data and simply pass on the maximum value which reduces the size of the output.

The ReLU function takes one input, x, and returns the maximum of {0, x}.

IV. EVALUATION METRICS

1) I want to use accuracy as evaluation metric for cell type classification.
2) Accuracy is a common metric for categorical classifiers.

$$\text{Accuracy} = \frac{\text{(images correctly classified)}}{\text{(all images)}}$$

3) The problem we have taken should contain high recall value.

a) *Recall:* Recall literally is how many of the *true* positives were *recalled*(found), i.e. how many of the correct hits were also found.

$$\text{Recall} = \frac{\text{True positive}}{\text{(True positive + False Negative)}}$$

b) *True Positives:* are the values which are correctly predicted as positives.
c) *True Negatives:* are the values which are correctly classified as negatives.
d) *False Positives:* are the values which are wrongly classified as positives. These are also type-1 errors.
e) *False Negatives:* are the values which are wrongly classified as negatives. These are also called as type-2 errors.

V. PREDICTIONS

Finding the accuracy of every model. Choose the best model out of ResNet50 and Standard CNN Model whose test accuracy will be more.

Models	Accuracy	Time
ResNet50 Model	95.6096%	1911 Seconds
Standard CNN Model	96.1901%	78 Seconds

The accuracy of Standard CNN model is high when compared to the ResNet50 model.

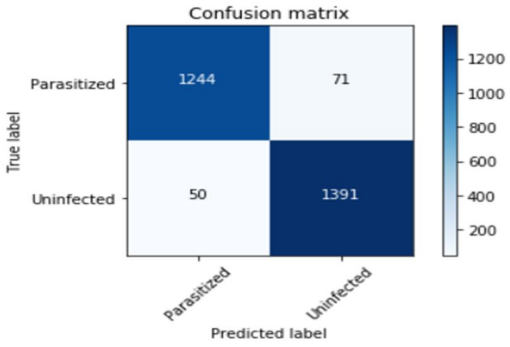
And the time take for Standard CNN model to train the entire data set is far less compared to ResNet50 model.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

VI. CONCLUSIONS

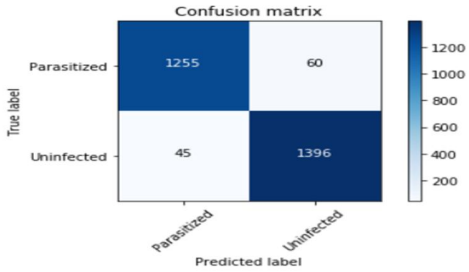
The conclusions obtained after plotting confusion matrices as

```
Confusion matrix
[[1244   71]
 [  50 1391]]
```



ResNet50 model

```
Confusion matrix
[[1255   60]
 [  45 1396]]
```



Standard CNN model

A. The misclassifications are observed more in ResNet50 model (121) than standard CNN model (105) which cause damage to human life.
B. Though Both models are giving almost same performances on training and testing data.
C. But we preferred standard CNN model as best, as its training time is far less than RESNET50 model.
D. As this is a high recall problem, standard CNN has high recall value than ResNet50.
E. So, we consider standard CNN model as the best.

VII. IMPROVEMENTS

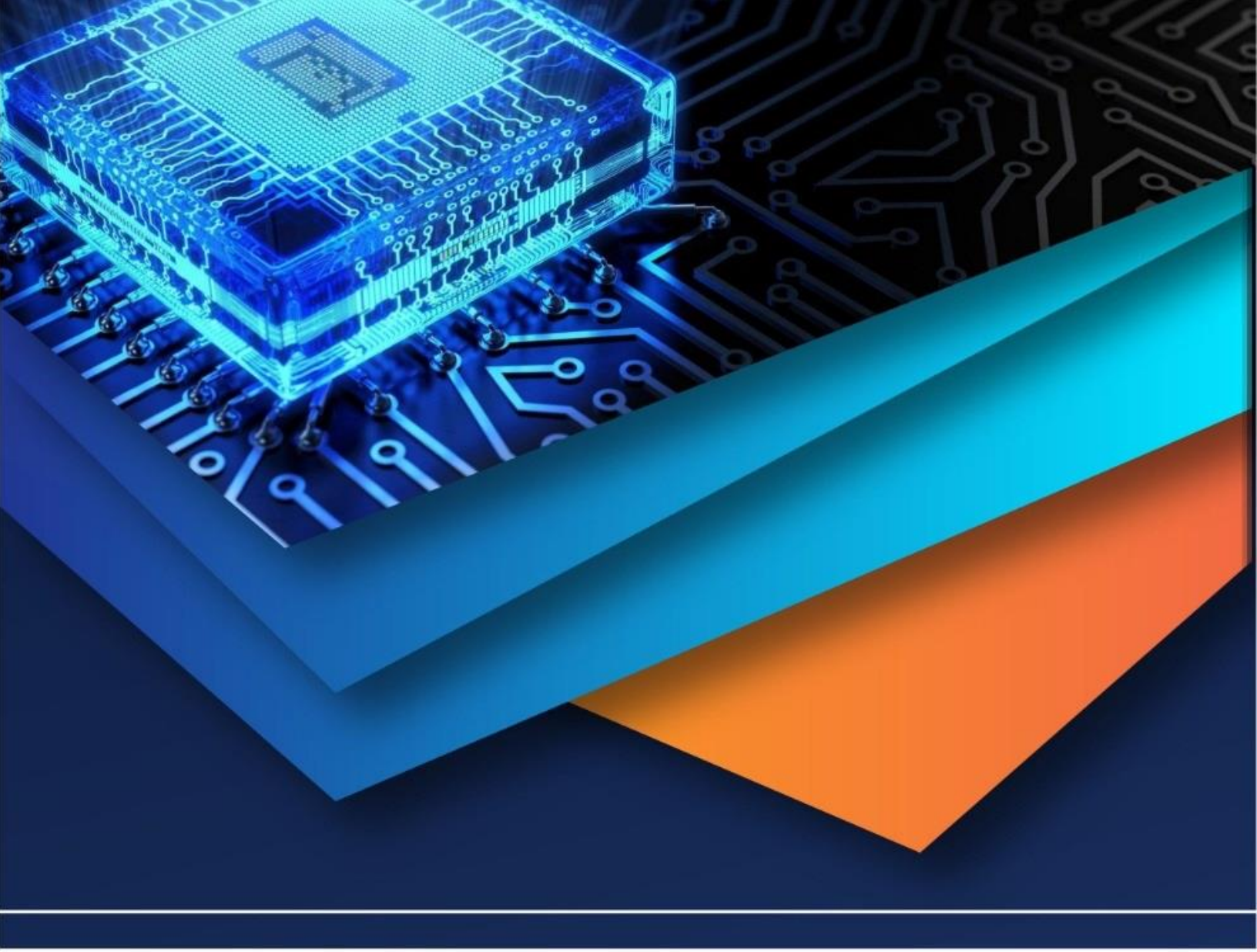
One thing, that can be improved from my models is we can use K-fold method for splitting data while fitting model, when we have enough time.

And also, use grid search for tuning hyper parameters but it takes so much time.

And also, data augmentation can be applied.

REFERENCES

[1] https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria/home
[2] https://ceb.nlm.nih.gov/repositories/malaria-datasets/
[3] https://unsplash.com/@ekamelev
[4] https://keras.rstudio.com/articles/sequential_model.html
[5] https://github.com/onnx/models/tree/master/resnet50
[6] https://keras.io/applications/#resnet



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY