DEVELOPMENT OF SUPERVISED AND UNSUPERVISED MACHINE LEARNING ALGORITHM FOR DETECTION OF MALARIA PARASITES IN THIN BLOOD SMEARS USING ORANGE SOFTWARE

BY

SULEIMAN, Jamila

(MTech/SPS/2018/8882)

SUBMITTED TO

THE DEPARTMENT OF PHYSICS,

SCHOOL OF PHYSICAL SCIENCES,

FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA NIGERIA

MARCH, 2023

DEVELOPMENT OF SUPERVISED AND UNSUPERVISED MACHINE LEARNING ALGORITHM FOR DETECTION OF MALARIA PARASITES IN THIN BLOOD SMEARS USING ORANGE SOFTWARE

BY

SULEIMAN, Jamila

(MTech/SPS/2018/8882)

SUBMITTED TO

A THESIS SUBMITTED TO THE POSTGRADUATE SCHOOL FEDERAL UNIVERSITY OF TECHNOLOGY MINNA, NIGERIA, IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF TECHNOLOGY (MTech) IN MEDICAL PHYSICS

MARCH, 2023

ABSTRACT

Highly sensitive and specific malaria diagnosis methods that are satisfactory for point-of-care testing in high burden areas are essential for productive treatment and monitoring of the disease. Microscopists often examine thick and thin blood smears which are the gold standard to diagnose malaria disease and compute parasitemia, Hence, the need for highly trained experts to interpret the data. In this study, machine learning algorithms for the detection of malaria parasite in thin blood smear images have been developed to reduce reliance on human proficiency, especially in the situations where experts are unavailable. The datasets containing 27558 cell images was obtained from National Library of Medicine, National Institute of Health (NIH) and used for both supervised and unsupervised machine learning models development. For supervised learning, logistic regression and random forest classifiers were used to predict the classes of thin blood smear images. These models classified the images as either uninfected or parasitised. Logistic regression returned a classification accuracy of 93.5% for parasitised images and 96.5% for uninfected smears. Random forest returned a classification accuracy of 90.5% for parasitised and 90.4% for uninfected smears. For unsupervised machine learning, hierarchical clustering and k-means models were implemented. Hierarchical clustering grouped parasitised images in one cluster and uninfected in another cluster and k-means gave a value of 0.218, discovered two clusters from the dataset. These results showed that logistic regression model produced the best performance for classification of thin blood smears of malaria. In cases where the classes of the smears are not known, the unsupervised machine learning models can be used to detect malaria infections in the smears. These models can be combined as backend programs for the design of a robust computerised malaria detection computer program. It is important to note that, although this method may not fully abolish the need for trained experts, the model implementations can be of great assistance in aiding the diagnostic decision-making process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

ANN	Artificial Neural Network
AUC	Area under Receiver-Operator Curve
CA	Classification Accuracy
CADx	Computer-aided Diagnostic
CDC	Centers for Disease Control and Prevention
CNN	Convolutional Neural Networks
DL	Deep Learning
FN	False Negatives
FP	False Positives
F1	F score
INLSVRC	ImageNet Large Scale Visual Recognition Challenge
kNN	K-Nearest Neighbours
LHNCBC	Lister Hill National Center for Biomedical Communications
ML	Machine Learning
MSE	Mean Squared Error
NIH	National Institute of Health
PCR	Polymerase Chain Reaction
RBCs	Red Blood Cells
RDTs	Rapid Diagnostic Tests
RGB	Red Green Blue
ROC	Receiver Operator Curve
ROI	Region of Interest
SVM	Support Vector Machine

TN	True Negatives
TP	True Positives
WHO	World Health Organization

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background to the Study

A female anopheles mosquito bite can spread the Plasmodium parasite, which causes malaria, a serious and potentially fatal disease. The red blood cells (RBCs) are infected by the parasites, which develop in the liver before being discharged directly into the blood and causing symptoms that can be fatal. Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale, Plasmodium Knowlesi, and Plasmodium malariae are among the parasite species that exist; however, Plasmodium falciparum can be fatal and affects the majority of the global population (WHO, 2018). The World Health Organization (WHO) latest report on global malaria estimates that there will be 241,000,000 instances of malaria and 627,000 fatalities from malaria worldwide in year 2020 (WHO, 2021). According to reports, children that are under five years of age are the most in danger; they make up 61% of the anticipated mortality tolls (WHO, 2018). Africa has the highest prevalence of the illness brought on by Plasmodium falciparum, South-East Asia and the Eastern Mediterranean come next (WHO, 2018). Malaria control and elimination methods have reportedly received a global investment of US $3.1 billion from disease-endemic nations (WHO, 2018). Malaria typically causes fever, exhaustion, headaches, as well as unconsciousness and convulsions in extreme situations, which can be fatal. The global case incidence rate for malaria is displayed in Figure 1.1. Sub-Saharan African nations and India bear a disproportionately heavy burden of the disease. Collectively, they are responsible for 85% of fatal cases. Among all fatalities, children under five years of age made up two thirds. A delay in diagnosis and treatment is one

of the leading causes of death in malaria patients. The most reliable and widely used method for illness diagnosis continues to be microscopic thick/thin-film blood analysis (CDC, 2018).

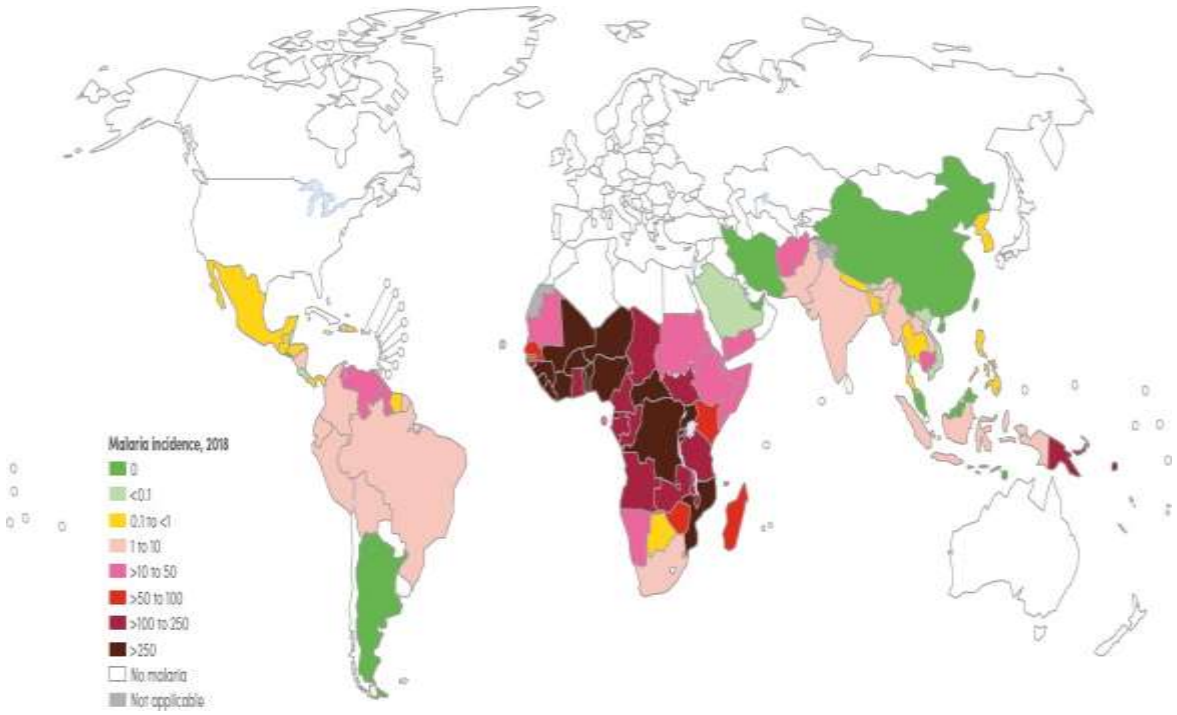


Figure 1.1: Map of malaria cases in 2018 (per 1000 population at risk) (WHO, 2019).

However, manual diagnosis is a laborious operation; the obligation imposed by elements like Inter- and intra-observer variation and widespread screening, especially in disease endemic nations in resource-limited circumstances, has a significant negative impact on the diagnostic accuracy (Mitiku *et al.*, 2003).

Risk assessment and medical diagnosis using images, computer-aided diagnostic (CADx) tools have become increasingly popular. These technologies analyze medical images for common manifestations and spotlights problematic abnormalities to support making medical decisions (Poostchi *et al.*, 2018). Nevertheless, most of these approaches to diagnosing

malaria employ manually created methods for feature extraction are tailored for specific datasets and instruction for variations in the region of interest (ROI)'s size, location, and orientation in the source machinery (Ross *et al.,* 2006).

By supporting triage and disease diagnosis, computer-aided diagnostic (CADx) tools incorporating machine learning (ML) algorithms on microscopic blood smear pictures significantly lessen the clinical burden (Poostchi *et al.,* 2018). By self-discovering the properties, data-driven deep learning (DL) approaches currently outperform handcrafted feature extraction methods when working with raw pixel data and performing end-to-end feature extraction and classification (LeCun et al., 2015). In particular, a family of DL models called convolutional neural networks (CNN) have demonstrated promising outcomes when classifying and recognizing images, and localization tasks (Redmon *et al.,* 2016).

1.2 Statement of the Research Problem

Exact parasite numbers are crucial for more than just diagnosing malaria. They are essential for determining the efficacy of medications, determining drug resistance, and categorizing disease severity (WHO, 2016). Microscopic diagnosis significantly is dependent on knowledge and expertise of the microscopist. In low-resource settings, microscopists usually labor by themselves because there isn't a strict system in place to assure their skill maintenance, which lowers the accuracy of diagnoses. This results in the field making inaccurate diagnostic conclusions (WHO, 2016). False negative results, or classifying an infected person as uninfected, result in the needless prescription of antibiotics, a second consultation, missed workdays, and in certain circumstances, the development of severe malaria as the illness progresses (Shillcutt *et al.,* 2008). False positive findings, or classifying an uninfected individual as infected, result in the inappropriate use of anti-malaria

medications and the possibility of experiencing side effects such as nausea, abdominal discomfort, diarrhea, and occasionally serious problems. Only 82% and 85%, respectively, are the best estimates for the sensitivity and specificity of microscopic diagnosis at district hospitals and health care centers in sub-Saharan countries (Shillcutt *et al.,* 2008). An attempt to perform malaria diagnosis automatically has been made in response to this sober study of the disease that is having sensitivity and specificity of less than 85%. When opposed to manual counting, automatic parasite counting offers the following benefits:

1. It gives blood films a more consistent and trustworthy interpretation.
2. By lessening the number of hours worked by malaria field workers, more patients can be served. Manual inspection is a challenging which takes time approach of identifying malaria and requires the pathologist's complete attention. Therefore, the creation of automated methods is essential for the quick and precise diagnosis of malaria. It can help in the early detection of disease so that it can be effectively treated and minimize the risk of false negatives (Mustafa *et al.,* 2021).
3. It can lower the price of diagnostics. Malaria parasites can be found using a variety of techniques. The automated parasite identification algorithm addresses the shortcomings of conventional approaches, such as high per-test costs, as compared to conventional diagnostic processes (Mustafa *et al.,* 2021).

1.3 Aim and Objectives of the Study

The aim of this study is to develop a supervised and unsupervised machine learning algorithms for detection of malaria parasite in thin blood smear images using Orange software.

The objectives of the study are to:

i. design k-nearest neighbours (kNN), support vector machine (SVM), random forest and logistic regression classifier algorithms for 27558 Giemsa-stained images.

ii. design and train a hierarchical clustering and k-means algorithms for classification of thin-smear Giemsa-stained images.

iii. evaluate model performance under supervised and unsupervised conditions.

1.4 Significance of the Study

Every year, millions of blood smear films are painstakingly inspected by skilled pathologists, and diagnosing malaria requires a significant human and financial investment. Additionally, reliable parasite counts from blood films are necessary for a proper diagnosis and grading of disease severity. If a patient did not have malaria cells but the doctor wrongly gave antibiotics, the patient would unnecessarily go through nausea or stomach pain (Grabias and Kumar, 2016). The ability to identify parasites throughout all stages of the malaria life cycle requires a robust malaria diagnosis with high sensitivity (fewer false negatives). Early, accurate diagnosis of malaria is fundamental in providing appropriate treatment and possibly reduce mortality rate.

1.5 Scope and Limitation of the Study

In this study, algorithms will be designed and trained using malaria datasets containing 27558 cell images with equal instances of parasitised and uninfected with malaria parasite obtained from National Library of Medicine, NIH. Lister Hill National Center for Biomedical Communications to detect malaria parasite in thin blood smears, this Giemsa-stained thin blood smears datasets are infected with Plasmodium falciparum. This is achieved using Orange software.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Malaria Infection

The protozoan parasites of the genus Plasmodium that cause malaria attack, red blood cells and spread through bites from infected female Anopheles mosquitoes (Poostch *et al.,* 2018). Specifically in Africa, where a child dies from malaria virtually each minute and where the disease is a major contributor to pediatric neurodisabilities, children die at a disproportionately high rate (WHO, 2016). The World Malaria Data 2016 estimates that 95 countries and territories, home to 3.2 billion people, are at risk of contracting malaria and becoming unwell, with over 1 billion of them at particularly high risk (more than one in one thousand probabilities of contracting the disease annually). In 2016, there were over 214 million cases of malaria worldwide, resulting in 438,000 fatalities. Africa bore the brunt of the load, accounting for 92% of all malaria deaths, according to estimates.

2.2 Malaria Causing Parasites

Human malaria is caused by five different Plasmodium species: Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale, and Plasmodium knowlesi. The two species that are most common are P. vivax and P. falciparum. Most deaths associated with malaria globally are caused by the most severe strain, P. falciparum. (WHO, 2016). In sub-Saharan Africa, P. falciparum is the most common malaria parasite and was thought to be responsible for 99% of all cases in 2016. The majority of malaria cases outside of Africa are caused by P. vivax, which is responsible for 64% of those instances in Americas, as well as over 30% in Southeast Asia and 40% in the Eastern Mediterranean (WHO, 2017).

Each of these parasite species undergoes phases during their growth cycle (which lasts 48 hours), giving the parasites a different visual appearance that may be observed under a microscope. In chronological order, these stages are the ring stage, trophozoite stage, schizont stage, and gametocyte stage. Figure 2.1 displays typical illustrations of every stage for every species. Most P. falciparum juvenile stage parasites are present in peripheral blood in non-severe malaria, however all stages may be present in severe malaria. Red blood cells infected with P. falciparum trophozoites are isolated from peripheral blood circulation by sticking to the capillary walls of critical organs. If the capillaries are blocked for newly infected cells by already infected cells, more advanced parasite stages (trophozoites and schizonts) will be visible in the peripheral circulation, which indicates a significant infection and a poor prognosis. (CDC, 2013).

2.3 Malaria Diagnosis

When traveling to regions where malaria is endemic, there are medications available to treat and even prevent infections. The sickness of malaria is curable. There is still no effective malaria vaccine, despite extensive research and field study in this area. When malaria is contracted, it spreads quickly, offers a serious risk of becoming severe and cerebral, and is often accompanied by neurologic symptoms brought on by P. falciparum infections. It is imperative to obtain a malaria diagnosis as soon as possible. Although there are several ways to detect malaria, there is still need for improvement in the cost, specificity, and ease of use the diagnostic assays that are currently available.

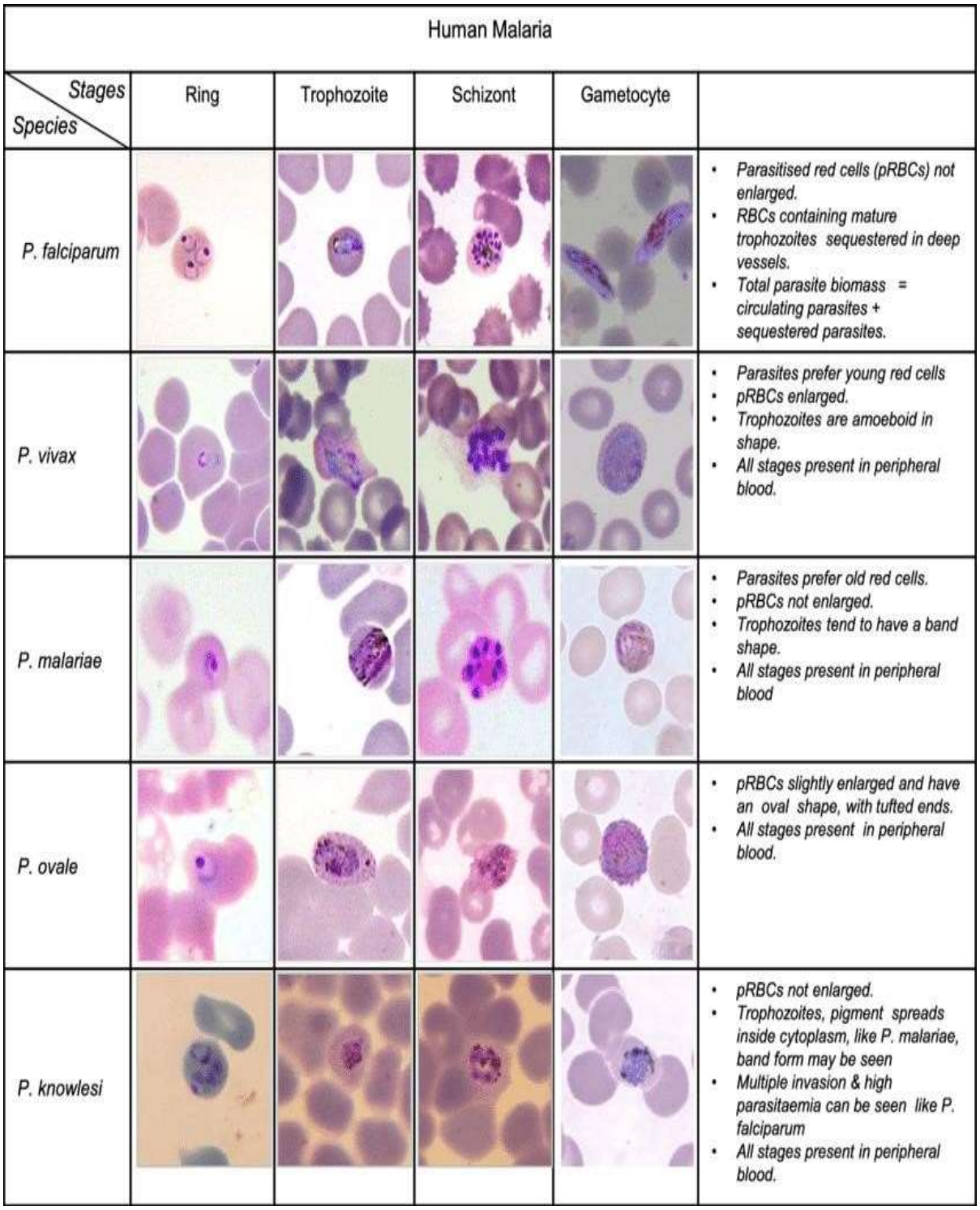


Figure 2.1: Five distinct Plasmodium species that cause human malaria and their various life phases in thin blood films (Silamut and White, 1993).

The first step in diagnosing malaria is to look for parasites. Identification of the parasite species, the existence of possibly combining infections, and watching of the stage of parasite development in connection to how bad the illness is are also crucial. Not only is counting parasites crucial for diagnosing infections and gauging their severity, but it also enables for patient monitoring by assessing therapeutic effectiveness as well as possible medication resistance.

2.3.1 Light microscopy

Although various methods of diagnosis exist and have recently gained popularity, light microscopy of blood films is the present industry-standard approach for diagnosing malaria in the field. Microscopy can be used to identify all parasite species since it can estimate the parasitemia level, clear a patient after a successful therapy, to monitor medication resistance. Additionally, it is more affordable and easily available than alternative methods. But its major limitations are the rigorous training required for a microscopist to become a skilled malaria slide reader, the expensive cost of both training and employment, along with the substantial amount of physical labor required. In order to diagnose malaria, a drop of the patient's blood is applied to a glass slide, which is then dipped in a staining solution to make parasites easier to see under a conventional light microscope, frequently with a 100 oil objective. Thin and thick blood smears are the two types of blood smears that are routinely prepared for the diagnosis of malaria (Poostchi *et al.,* 2018).

A thick smear is necessary to identify parasites in a drop of blood. Thick smears, which have an 11 times higher sensitivity than thin smears, allow for a more accurate detection of parasites. The blood drop is dispersed throughout the glass slide; however, this results in

narrow streaks that have additional advantages. They make it simpler for the examiner to recognize different malaria types and different stages of the parasite (Jan *et al.,* 2018).

2.3.1.1 The Physics of Light Microscope

The eye is amazing at seeing objects both big and small, but it is plainly restricted in the tiniest details it can pick up. The use of optical devices was motivated by the desire to see beyond the limit of the naked eye could see. The light microscope is a device for seeing an object's tiny features. It achieves this by using a sequence of glass lenses to first focus a light beam onto or through an object, then convex objective lenses to expand the image created. A straightforward convex lens can magnify an image, but it is challenging to achieve high magnification with such a lens. It is challenging to magnify an image by more than 5 without the image becoming distorted. We can add one or more extra lenses to the basic magnifying glass to achieve a higher magnification.

In the Netherlands and Denmark, eyeglass manufacturers invented the first microscopes in the early 1600s. Figure 2.2 illustrates the construction of the most basic compound microscope, which consists of two convex lenses. With a typical magnification range of 5 to 100, the objective lens is a convex lens with a short focal length and high power. A convex lens with a larger focal length is used in the eyepiece, which is also known as the ocular. The goal of a microscope is to enlarge little objects, and both lenses work together to achieve this goal. The eye cannot focus on objects or images that are too close, the final enlarged image is also produced adequately distance from the viewer to be easily perceived (closer than the near point of the eye). An objective and an eyepiece are the two lenses that make up a compound microscope. The object which is larger in the first image, is formed by the objective. The

eyepiece's focal length is occupied by the first image, which also acts as the eyepiece's target. The eyepiece creates the final, enlarged image (Ling *et al.*, 2016).

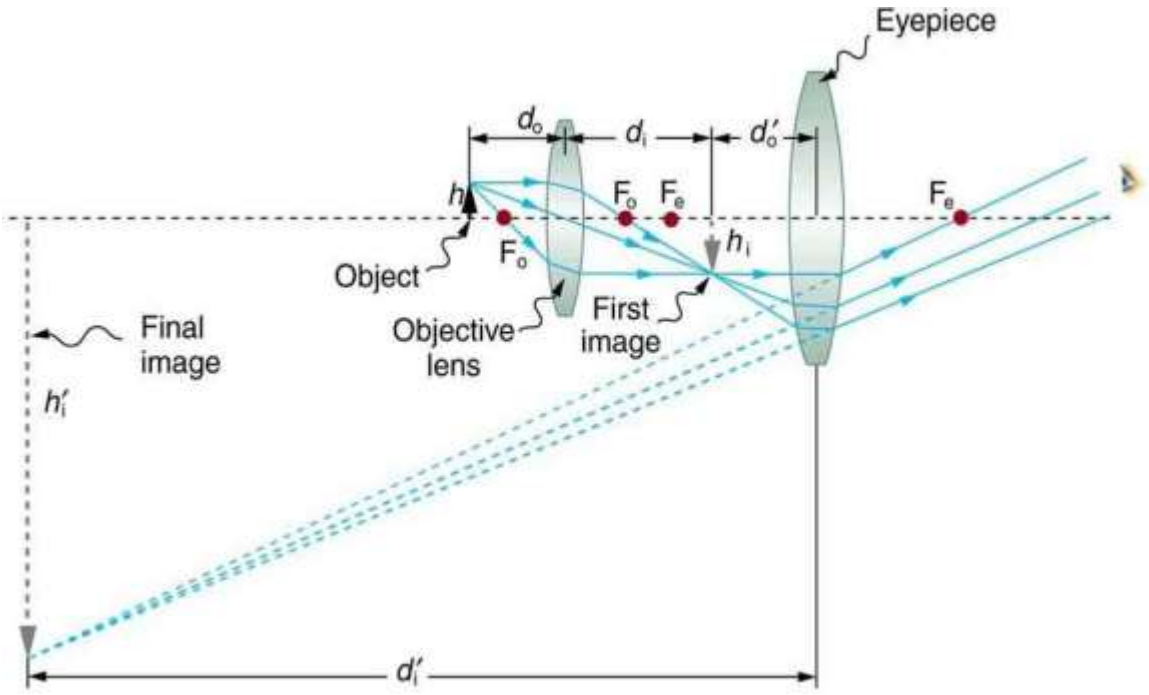


Figure 2.2: A compound microscope with two lenses, an objective, and an eyepiece. (Ling *et al.*, 2016)

Consider the two lenses on the microscope in Figure 2.2 in turn to observe how a picture is created. The objective lens's focal length f^{obj} is just past the object, creating a true, inverted image that is bigger than the actual thing. The subject of the second lens or eyepiece forms initial image. The first image is placed within the eyepiece's focal length f^{eye}, so that it can be further magnified. It therefore enhances the intermediate image created by the objective, like a magnifying glass. A magnified virtual image is what the eyepiece creates. The resulting image is still inverted but is visible since it is farther away from the viewer than the object.

The eye sees a virtual image that is projected by the eyepiece, which acts as the object for the eye's lens. Because the virtual image produced by the eyepiece is far outside of the eye's focus length, the eye produces a real image on the retina.

The linear magnification m^{obj} by the objective and the angular magnification M^{eye} by the eyepiece combine to provide the microscope's overall magnification. These are respectively given in Equation 2.1 and 2.2:

$$m^{obj} = \frac{d_i^{obj}}{d_o^{obj}} \approx -\frac{d_i^{obj}}{f^{obj}} (linear\ magnification\ by\ objective) \quad (2.1)$$

$$M^{eye} = 1 + \frac{25cm}{f^{eye}} (angular\ magnification\ by\ eyepiece) \quad (2.2)$$

Here, the focal lengths of the objective and eyepiece are, respectively, $f^{obj}$and f^{eye} We assume that the near spot of the eye, which provides the greatest magnification, is where the final image is generated. The eyepiece's angular magnification is the same as that of a standard magnifying glass. This shouldn't come as a surprise because the eyepiece functions similarly to a magnifying glass in terms of physics. The compound microscope's net magnification M_{net} is the sum of the angular and linear magnifications of the eyepiece and objective, respectively, the net magnification is given in Equation 2.3:

$$M_{net} = m^{obj} M^{eye} = \frac{d_i^{obj}(f^{eye} + 25cm)}{f^{obj} f^{eye}} \quad (2.3)$$

2.3.2 Rapid diagnostic tests

The fundamental benefit of microscopic malaria diagnosis is its cheap direct cost, which makes it stand out in environments with limited resources (WHO, 2016). Given the limited financial resources that are often available in regions where malaria is prevalent, other diagnostic techniques that are currently in use as well as any new techniques must demonstrate that they can offer the same simplicity of use and affordability as microscopy. Rapid diagnostic tests (RDTs) may be the sole and primary rival in this regard. They take around 10-15 minutes to process and look for antibodies, which are the parasites' telltale signs. They do not need any special equipment and merely need minimal training. Their detection responsiveness is lower but similar to manual microscopy. However, in high-burden locations, RDTs are now more expensive than microscopy (WHO, 2018). RDTs are utilized more commonly in rural areas without access to microscopy. RDT was used to conduct roughly 47% of tests for malaria in nations where the disease is endemic (WHO, 2016).

2.3.3 Other malaria tests

There are numerous ways to diagnose malaria. Prices for tests, as well as their sensitivity and specificity, duration per test, and the necessary user level competence are crucial factors. Additionally, counting the amount of infected red blood cells is crucial as a prognostic sign (Vink *et al.*, 2013).

i. Polymerase chain reaction (PCR): Compared to traditional microscopic inspection of stained peripheral blood smears, PCR has demonstrated improved sensitivity and specificity. In fact, it is thought to be the most accurate test out of all of them. It can distinguish between different species and very low parasite detection quantities in blood. However, PCR is a time-consuming, expensive, and sophisticated technology

that requires skilled personnel to process. The difficulty due to testing and the lack of resources to carry out these tests properly and regularly, according to Tangpukdee *et al*. (2009), are the main reasons that PCR is not commonly used in developing nations. The PCR method also requires quality assurance and equipment upkeep; therefore, it might not be appropriate for determining the presence of malaria in remote rural areas or even in standard clinical diagnostic settings.

ii. Fluorescent microscopy: A laboratory test called quantitative buffy coat uses fluorescence microscopy to find blood parasites like malaria. Parasites are visible under UV light thanks to a fluorescent dye. Adeoye and Nga (2007) claim that This test entails more accuracy than the typical thick smear. Commercially available fluorescent dye-infused portable microscopes used to identify parasites are now available. Nevertheless, the quantitative buffy coat method is straightforward, dependable, and user-friendly. It is less effective at identifying the types and numbers of parasites and necessitates specialized equipment that is more expensive than traditional light microscopy (Tangpukdee *et al.*, 2009).

iii. Flow cytometry: This technique for counting and detecting cells uses lasers and can profile hundreds of cells each second. Although automated parasitemia counts are available with flow cytometry, the sensitivity is very limited. When a clear answer is needed to make treatment decisions, In the real world, flow cytometry is less useful as a diagnostic technique. However, it can be used in a therapeutic environment in affluent nations to accurately measure the number of parasites, for example, in the follow-up of pharmacological therapy (Janse and Van Vianen, 1994).

2.4 Staining Methods

Giemsa's stain was used for the first time to diagnose malaria in 1902, more than a century ago. Since then, it has drawn further attention. It is being utilized often in microscopical malaria investigations because to its inexpensive cost, excellent sensitivity, and specificity (Keister *et al.*, 2002). However, Giemsa staining is labour-intensive, time-consuming, and requires many chemicals and experienced personnel (it typically requires at least 45 minutes to stain a slide).

Other stains have also been used, such as Field stain, which considerably shortens staining time but necessitates drying samples both before and during staining (Houwen, 2002). Field's stain can have drawbacks, particularly in health facilities with limited resources where it might be applied. Poor blood preparation frequently produces artifacts like bacteria, fungi, stain precipitation, dirt, and cell debris that are frequently misinterpreted for malaria parasites. False-positive readings can commonly be brought on by these.

High sensitivity Leishman's stain was discovered in 1901. It is affordable, and rather simple to use. One of the other stains in use is the Wright-Giemsa stain, which combines the Wright and Giemsa stains and allows for easier distinction of different blood cell types. Shute and Sodeman (1973) looked at the utility of fluorochrome staining for detecting malaria parasites in low-infection samples in the 1970s. Romanowsky and Giemsa staining techniques have been demonstrated to be less accurate and time-consuming than fluorochrome staining (Suwalka *et al.*, 2012). It has drawbacks such photo bleaching and phototoxicity, as well as requiring a lot of work and training. Additionally, the cost of fluorescence microscopes is more than a conventional light microscope, which is a problem in tropical areas with limited resources and high rates of malaria (Kawamoto *et al.*, 1971).

2.5 Introduction to Machine Learning

Machine learning, one of the most promising and rapidly expanding fields of computer technology, is the study of algorithms that enhance the effectiveness of machines or computers automatically by the training and testing of the machine or computers with undoubtedly varied variables (Smola *et al.,* 2008). Machine learning is the process of teaching a computer to use various algorithms to process data intelligently and automatically. Machine learning improves the accuracy and efficiency of data processing and is utilized in many different industries. Effective algorithms are used to create machine learning, which uses a specific collection of tools and functions to handle complicated and massive data problems. Machine learning is assisting in a wide range of industries. These applications of artificial intelligence are typically utilized for recognition and prediction in fields like computer engineering and medicine. Machine learning has reduced manual work for those who might be prone to mistakes and inaccuracy (Smola *et al.,* 2008).

Machine learning techniques can either be supervised or unsupervised, despite the fact that some authors also refer to other algorithms as reinforcement learning since they learn data and uncover patterns in order to respond to an environment. But the majority of publications acknowledge both supervised and unsupervised machine learning techniques. These two main classes are distinguished by labels present in the training data subset (Kotsiantis, 2007).

2.5.1 Supervised learning

A set of paired input-output training samples are used in the supervised learning machine learning paradigm to understand the details of a system's input-output connection. Because the output is regarded as the label of the input data or the supervision, an input-output training

sample is sometimes referred to as labelled training data or supervised data. On occasion, it may also be referred to as "Learning with a Teacher" (Haykin, 1998), The input-output connection knowledge of a system is learned using a collection of paired input-output training samples in the supervised learning machine learning paradigm.

Learning from Labeled Data or Inductive Machine Learning (Kotsiantis, 2007). The goal of supervised learning is to create an artificial system that can understand the relationship between the input and the output and predict the system's output given new inputs. If the output accepts a finite number of discrete values that represent the class labels of the input, the learned mapping categorizes the incoming data. The input is regressed as a result if the output is continuous.

2.5.2 Unsupervised learning

Unsupervised learning makes use of training data that are not labeled, classed, or categorized (also known as knowledge discovery). Unsupervised learning's main objective is to explore unlabeled data for intriguing and hidden patterns. Unsupervised learning techniques, in contrast to supervised learning, cannot be used to solve a regression or classification problem directly because it is unknown what the output values will be. The most popular unsupervised learning approach for exploring data analysis to uncover hidden patterns or groupings in the data is clustering (El Bouchefry and de Souza, 2020). Applications for cluster analysis include market research, object identification, and DNA sequence analysis. neural networks, clustering, anomaly detection and methods for learning latent variable models are typical unsupervised learning techniques (El Bouchefry and de Souza, 2020).

2.6 Conventional Malaria Image Analysis Techniques

Most algorithms suggested in the literature are centered on the categorization of thin-smear Giemsa-stained pictures obtained using stained blood smears during light microscopy. They frequently take the following steps to accomplish their goal of automatically counting all uninfected and parasitized erythrocytes: (1) pre-processing the blood smear image; (2) segmenting the erythrocytes from the background; (3) extracting parasite features; and (4) categorizing the erythrocytes mathematically. Figure 2.3 provides a graphic representation of this strategy. Below are examples of the methods utilized for each phase.

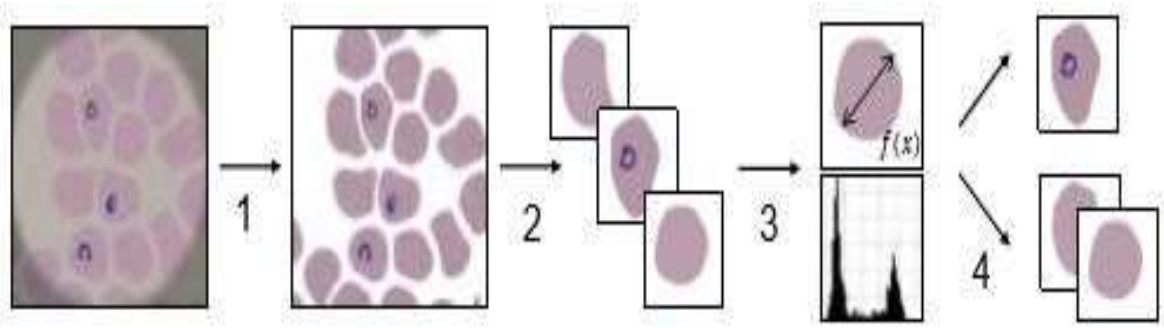


Figure 2.3: Schematic illustration of the fundamental image processing process used by the majority of (conventional) automated systems for diagnosing malaria. (van Driel, 2020).

1. Pre-processing

Pre-processing is frequently the initial step when undertaking digital analysis on any type of image data with the goal of reducing noise and improving image quality. There are many well-known filters for noise removal, including median and Gaussian. Each pixel value in median filters is simply replaced by the median of pixels in a radius around it. In Gaussian filters, a neighbourhood-weighted average of each pixel is calculated using a Gaussian distribution function in two dimensions, and that value is then substituted for the original pixel value. Though more complex filtering techniques have also been used, these

fundamental filters are frequently used in proposed automated malaria detection systems because they effectively eliminate noise (Linder *et al.*, 2014).

Low contrast is another frequent issue that is typically resolved using techniques like contrast stretching or histogram equalization. Contrast stretching is a linear normalization that enlarges the goal interval for an image's intensity range. The non-linear normalization known as "histogram equalization" enlarges the histogram regions where intensities are concentrated and shrinks the regions with low abundance intensities (Nasir *et al.*, 2012). Uneven lighting and differences in staining colour are additional issues common to Giemsa stained thick and thin film microscopic pictures. Gray world assumption is one of the colour normalization methods that can be used to fix this (Lam., 2005).

2. Segmentation

Segmenting the individual erythrocytes is quite simple when the thin smear is of acceptable quality, The image is sharp and well-lit, with important cells completely separated. It can be done using simple thresholding methods, such as Otsu's, which separates pixel values into two bins in an ideal way. When the image is highly bimodal, which is partially obtained through pre-processing, this works well (Anggraini *et al.*, 2011).

K-means clustering is an excellent substitute to iteratively assign pixels to foreground or background when bimodality cannot be obtained through pre-processing or when the image is blurred. Its drawback is that thresholding approaches are less computationally complex (Savkare *et al.*, 2015). When cells are contacting or overlapping, both approaches have issues. Many techniques have been suggested to separate individual erythrocytes when this is the case. Some simply thresholds the larger things repeatedly until only those that are roughly the right size are left. Under certain conditions, this approach can be effective, but it is not very

reliable (Mushabe *et al.,* 2013). Another well-liked cell segmentation approach is water shedding, however for it to work, the objects' boundary gradients must not be too weak (Sharif *et al.,* 2012). Circle Hough Transforms have also been utilized and can be effective, but they err when erythrocytes wander too much from their fixed size and circular shape assumptions (Zou *et al.,* 2010).

3. Feature extraction

The term "feature extraction" in pattern recognition refers to the process of calculating values from the raw data (pixel) data that will best give information for the classification you wish to do, without information loss or duplication. Colour values of pixels are evidently informative aspects for diagnosing infection in blood slides containing stained parasites. These can be used to compute features including co-occurrence matrices, local binary patterns, and histograms of oriented gradients. Given that there is the greatest contrast between the stained parasite and the erythrocyte within the green channel of an image in RGB color space, some authors have expressly suggested solely collecting color data from this channel. Others have recommended utilizing a combination of the two or converting the image to HSB-space before extracting the color features. To assist in classification, morphological features can also compute measurements like relative form measures and granulometry (Devi *et al.,* 2016).

4. Classification

When dividing objects over classes, the objective is to minimise Interclass disparity, based on the object features supplied. Essentially, an approximate mapping f from the input features x to the output class y is produced by a classification algorithm, such that $f(x) \approx$ y. An example of a simple classification method is the earlier mentioned ‘thresholding’, where objects are divided into classes based on whether its value falls or rises above a specific level. More

complicated classification methods often use a training set of previously categorized objects to discover a classification method with the lowest error rate, which is called ‘supervised learning’.

‘Unsupervised learning’, where all that is known are the input data and the cost function. a priori, is also possible. A great number of learning algorithms have been developed, such as Support Vector Machine (SVM), Bayesian classifiers, K-nearest neighbour classifiers, logistic regression trees, artificial neural networks, among other things. These have all been used in the analysis of thin smears containing malaria. A typical goal is binary classification, which divides objects into parasitized and parasitic erythrocytes. However, attempts have also been made to further categorize parasitized cells into 20 types (Tek *et al.*, 2010). The standard and distinguishability of the features that were retrieved from the parasites and erythrocytes are key factors in the efficacy of these approaches.

2.7 Neural Networks and Malaria Image Analysis

There is currently no standardised comparison for malaria picture classification, therefore identifying the state-of-the-art is incredibly challenging. Such comparisons are however feasible in more general picture classification research. It is abundantly obvious from image classification competitions like the ImageNet Large Scale Visual Recognition Challenge (INLSVRC) that Artificial Neural Network (ANN)-based deep learning techniques have taken over the field in recent years (Krizhevsky *et al.*, 2017).

2.8 Mathematical Principles

2.8.1 Basic principle

An ANN is a classifier that combines feature extraction and classification into a single algorithm. It was inspired by biological neural networks. The most basic kind of ANN is a feed forward neural network, commonly referred to as a multilayer perceptron.

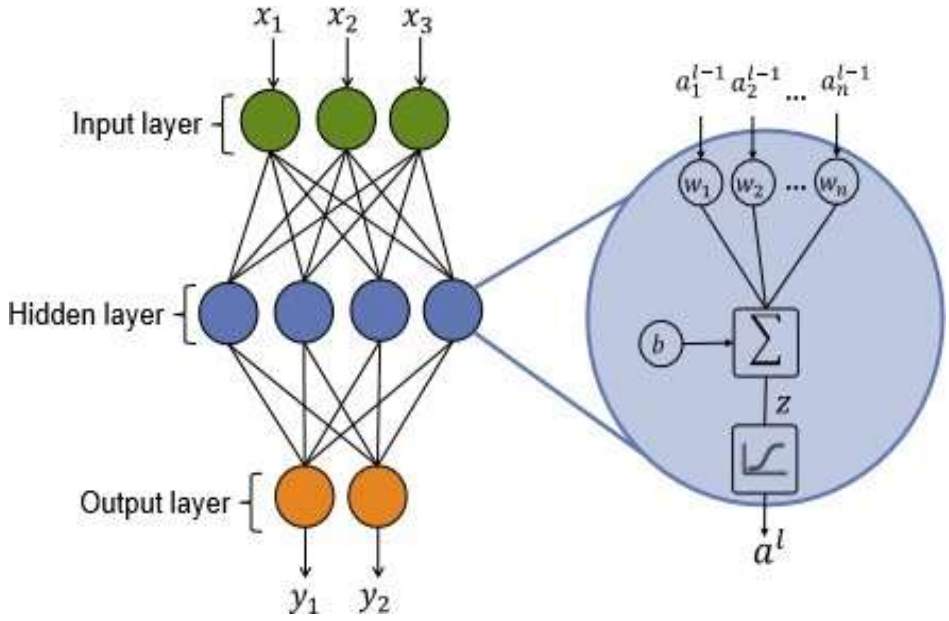


Figure 2.4: A feed-forward neural network with three inputs, two outputs, and one hidden layer is shown schematically (van Driel, 2020).

They consist of an input layer that contains all the data input points, an output layer that maps inputs to outputs, and (optionally) any number of hidden layers. The ANN is said to be "completely linked" if all nodes in all layer's pass outputs to one another, as shown in figure 2.4. Deep neural networks are frequently referred to as such when numerous hidden layers are incorporated into the network's architecture, and deep learning is the term used to describe both the network's training and use. Artificial neurons make up the buried layers. The inputs are transformed in each of these neurons using a non-linear activation function in conjunction

with an affine transformation. Let the result of a single neuron k in layer l, be denoted $a^l{}_k$. Each neuron uses the vector of outputs of the previous layer a^{l-1} as inputs, the first step is to compute a weighted sum z^l_k of these as given in Equation 2.4

$$z^l_k = \sum_{i=1}^{n} (a_i^{l-1} w^l_{ki}) \quad (2.4)$$

where n denotes the dimension of the preceding layer w_{k1} ...w_{kn} are weights of the neuron. A bias b_k is added, and the output is then computed by applying some non-linear activation function g given in Equation 2.5

$$a^l_k = g(z^l_k + b^l_k) \quad (2.5)$$

This output is then propagated to the neurons in the next layer, where they the same type of transformation. The total mapping of the inputs x to outputs y is thus a function of all the weights and biases; $\hat{y}$ = f(x,W,b). The correct mapping from the inputs to the outputs is approximated such that $\hat{y} \approx$ y by adjusting the weights and biases during learning. Neural networks have demonstrated their universality function approximators, meaning that any mapping can be approximated arbitrarily well, given enough hidden units are used (Chen and Chen, 1995).

2.8.2 Convolutional layers

A Convolutional Neural Network (CNN) is a type of deep neural network that was developed specifically with the aim of image classification. A core concept in the architecture of CNNs is the introduction of convolutional layers. The input of each neuron is a function of only a small region of the outputs of the previous layer. This input is produced by convolving the previous layer with a small matrix of weight called the kernel. The kernel slides over the original image and the convolution of the kernel with the region surrounding the input pixel is computed by as given in Equation 2.6

$$z_{ij} = W * x_{ij} = \sum_a \sum_b w_{ab} \quad . x_{(i-a)(j-b)} \tag{2.6}$$

where $w_{ab} \in w_{00} \ldots w_{NN}$ are the weights in the kernel W of size $N \times N$ and $x_{ij} \in x_{00} \ldots x_{nn}$ are the values of the input matrix X with size $n \times n$. The convolution z_{ij} is then used to activate a function, which results in the output y_{ij} as shown in Equation 2.7:

$$y_{ij} = g(z_{ij} + b) \tag{2.7}$$

Equations 2.6 and 2.7 replace equations 2.4 and 2.5. Besides this, the convolutional layers are implemented in the same way as the standard network layers described above. The neurons in convolutional layers are structured in a grid, this makes convolutional layer especially suitable for categorizing structured data, such image data. The kernel essentially acts as a feature extraction filter, where the learnable weights converge towards features in the image. By using the same kernel with the same weights on the entirety of the input, an activation map of these features is produced. Therefore, a feature map is the name given to the convolutional layer's output. The convolutional layer operation is shown schematically in figure 2.5

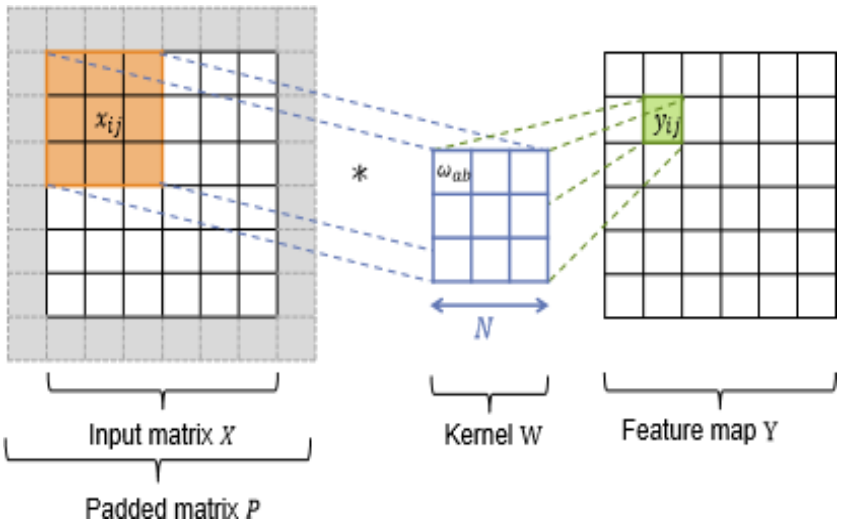


Figure 2.5: Convolution of a 3x3 kernel and a 6x6 input picture is shown schematically. Padding is utilized to create a 6x6 feature map (van Driel, 2020).

Given an $n \times n$ image X as input, and a $N \times N$ kernel W, which slides over the input matrix with stride 1 (meaning it moves 1 pixel for each convolution), the size of the feature map will be $n - N + 1 \times n - N + 1$. When a feature map of equal size to the input is desired, padding can be used around the input matrix. This is also depicted in figure 2.5. Often, multiple kernels are used in one convolutional layer to produce multiple feature maps. If M kernels are used, the size of the output (with padding) will be $n \times n \times M$. The convolution described above assumes a single channel input. It is possible to have a multi-channel input to a convolutional layer. In this case, the convolution can be described as in Equation 2.8:

$$z_{ij} = W * x_{ij}^{k} = \sum_{k=1}^{K} (\sum_{a} \sum_{b} w_{ab}^{k} . x_{(i-a)(j-b)}^{k} \quad (2.8)$$

Here, x_{ij}^{k} refer to the pixels in the kth input channel, the total the quantity of input channels is K. The kernel in this case takes the size $N \times N \times K$, but the output remains two dimensional. Even though the kernel is now 3D, this is still referred to as a 2D convolution, mainly because the kernel obstructs the input only in horizontal and vertical direction. It can be thought of as a stack of filters, where each filter is convolved with one input channel, and the outputs of the convolution are summed.

The dimensionality should be minimized to avoid over-fitting in CNNs, pooling layers are often added after convolutional layers. In these, the convolutional layer outputs are downsampled. The $n \times n$ feature map is reduced in size to $\frac{n}{p} \times \frac{n}{p}$, by dividing the feature map in $p \times p$ patches and taking some function of the values in this patch as the output. In average pooling layers, the average of the values is passed, while Max pooling layers pass the largest value. It is also possible to up-sample through convolution, when a feature map of a bigger size than the input is desired. This concept was introduced as ‘deconvolution’, but

'transposed convolution' has since been suggested to be a more accurate name (Zeiler *et al.,* 2010).

2.8.3 Model training

The first step in training the ANN is initialisation. In order to ensure convergence of the network it is important that the outputs of layers don't explode or vanish after the first pass. Initialising the weights and biases in such a way that the standard deviation of the activation outputs of each layer is normalized is a good way to prevent this. In order to achieve this, the 'Xavier initialisation' was proposed, where the weights of a layer are drawn from a uniform set, which is bounded between $\pm\frac{\sqrt{6}}{\sqrt{n_i+n_{i+1}}}$, where n refers to the number of incoming network connections, and n_{i+1} the number of outgoing connections (Glorot and Bengio, 2010). This strategy works well for continuous activation functions that are symmetric about zero, such as tanh. For asymmetric functions such as ReLU, an initialization dubbed the 'Kaiming initialization', in which weights are randomly drawn from a standard normal distribution and scaled by $\frac{\sqrt{2}}{\sqrt{n_i}}$ was shown to lead to faster convergence (Dong *et al.,* 2017).

Each training iteration of the network can be divided into three phases: forward propagation of the data, backward propagation and optimisation. During forward propagation, the prediction $\hat{y}$ of the current network on the data is computed, by computing equation 2.5 for every neuron in every layer. This prediction is used to determine the loss function's value J, which is some measure of the total error in the system. Often the Mean Squared Error (MSE) given in Equation 2.9 is used;

$$J = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2.9)$$

where y is a vector containing the ground truth of the network. Other options for loss functions that are commonly used are the Root Mean Squared Error and the Mean Absolute Error. These loss functions are effective when the targeted output is a continuous value. When dealing with classification, the target output is one of integer classes. In this case, cross entropy is a more effective measure of the error in the system, and therefore often used as loss-function. When dealing with a two class classification problem, the binary cross-entropy loss function is given by Equation 2.10:

$$J = -\frac{1}{n}\sum_{i=1}^{n} y_i . \log(\hat{y}_i) + (1 - y_i). \log(1 - \hat{y}_i) \tag{2.10}$$

It is possible to add additional terms to the loss function to influence the outcome, for example, a regularization term that penalizes large weights, which can help reduce over-fitting can be added. When extra terms are added, the objective function is no longer only a function of the loss and is therefore referred to as cost function. The next phase is back propagation, during which the gradient of the cost function is calculated. This is done by computing an error function δ^l at each layer, by calculating the cost function's derivative with respect to the weighted inputs z^l:

$$\delta^l = \frac{\partial J}{\partial z_l} = \sum_k \frac{\partial J}{\partial a_k^l} \frac{\partial a_k^l}{\partial z_k^l} = \nabla_a J \circ g'(z^l) \tag{2.11}$$

where $\nabla_a J$ is a vector of derivatives of J with respect to the components of a^l. In the output layer L, these components are known ($\hat{y} = a^L$), making it easy to compute the error of the output layer. The error of each neuron can then be calculated by propagating this back through the network. An equation for the error at a layer $l - 1$, in terms of its succeeding layer is given by Equation 2.12

$$\delta^{l-1} = ((W^l)^T \delta^l) \circ g'(z^{l-1}) \tag{2.12}$$

This can be used to compute the errors all the way through the network efficiently. When errors are known, these can be used to compute the gradient of the network by realizing that Equation 2.13:

$$\frac{\partial J}{\partial b_k^l} = \delta^l; \qquad \frac{\partial J}{\partial w_{ki}^l} = a_k^{l-1}\delta^l \tag{2.13}$$

The gradient is finally used to update the weights and biases during optimization. A gradient descent method is used for this; since the gradient gives the direction of the largest increase of the cost function, to minimize it, a step in the opposite direction of the gradient is taken as given in Equation 2.14:

$$w_{ki}^l \leftarrow w_{ki}^l - \alpha \frac{\partial}{\partial w_{ki}^l} J; \qquad b_k^l \leftarrow b_k^l - \alpha \frac{\partial J}{\partial b_k^l} J \tag{2.14}$$

This step's dimensions α is called the 'learning rate' and it is a tuneable parameter in training the network. Often, the training samples are divided into batches, and the gradient is determined for all training samples in the batch before updating the weights. The size of this batch is a hyper-parameter of the network which can be tuned to achieve the desired performance. Small batch size leads to stochastic weight updates, while large batch size leads to slow learning. An epoch is defined as the number of iterations after which all training data has been passed through the network exactly once.

2.9 Review of Related Works

The use of neural networks to categorize Giemsa-stained, malaria-infected blood smears has been studied in some publications.

Savkare *et al*. (2011) gathered Red, Blue, Green (RGB) images and performed typical histogram equalization, median, and Laplacian pre-processing. They made the image grayscale, applied Otsu thresholding on the grayscale and green channel, and then blended the two independent binary masks into one. An average erythrocyte size was determined, and

items that did not fit this standard were deleted as artifacts or leukocytes, leaving behind a binary mask of background and erythrocyte objects. Erythrocyte recognition was reported to have a success rate of 99.43%; however, in calculating this number, objects made up of several erythrocytes were considered correctly recognized. These items were divided using water shedding, and any fragments that were too small to be erythrocytes were eliminated.

The eliminated objects were probably the consequence of over-segmentation given the high accuracy recorded before to the water shedding, but as no separate accuracy is reported following this step, it is hard to determine how many split erythrocytes were successfully discovered using this technique. The third moment, the mean, and the standard deviation of the green channel histogram for the produced objects were computed, along with their shape and textural characteristics. These were used to employ an SVM to determine whether erythrocytes were infected; the classification's stated sensitivity and specificity were 93.12% and 93.17%, respectively.

Plasmodium vivax and Plasmodium falciparum-infected thin blood smears were studied by Das et al. (2013) to determine their classification. Prior to applying marker-controlled water shedding to segment erythrocytes, they used gray world assumption to correct illumination and a geometric mean filter to eliminate noise from their images. There were no specific performance metrics supplied for the segmentation. The most important features were chosen by statistical analysis after they computed 96 textural and morphological features altogether. After that, erythrocytes were classified as infected or non-infected as well as distinguishing between the 5P. vivax and P. falciparum life stages using a Bayesian classifier and an SVM.

The Bayesian classifier was able to complete this task with accuracy of 84% (Das *et al.,* 2013).

A modest dataset of segmented erythrocytes was used, objects were collected by thresholding and then applying a Hough circle transform to blood slide pictures, Dong et al. (2017) trained three distinct CNN architectures. They did this to produce equal-sized training and testing sets, each containing 517 infected cells and 765 uninfected cells, respectively. The segmentation's performance data were not provided. On these photos, the LeNet-5, AlexNet, and GoogLeNet architectures were trained, and accuracy results were reported for each network as 96.18%, 95.97%, and 98.17%, respectively. This was contrasted with an SVM that was trained using the same data and methods as (Das *et al.*, 2013) which had an accuracy of 91.66%.

The classification of Giemsa-stained thin films was another area of research for Rajamaran *et al*. (2018). They started by using a standard cell segmentation technique to separate the erythrocytes from blood slide images.

They compiled a collection of 27,558 images of cells, equally split between those with parasites and those without them, made it available to the public, and then created a CNN-based classifier for it.

They suggested a three-block, two-convolutional layer network architecture, with the first block having a max pooling layer, the second having an average pooling layer, and the third block having three fully linked layers straight after. They attained sensitivity and specificity of 93.11 and 95.12 on the object level. Later, the effectiveness of their suggested network

architecture was compared with that of already-used network architectures like VGG-16 and ResNet-50, and these performed somewhat better.

Instead of using just one image per cell object, Gopakumar *et al*. (2018) suggested training a network on a focus stack of RGB cell images. It was asserted that this would enhance performance in separating parasites from artifacts like dust grains. A two-stage threshold-based acquisition technique was used to acquire segmented cells. There were no specifics given regarding the CNN's architecture. There were reported values for sensitivity and specificity of 96.98% and 98.50%, respectively. However, at 173% of the actual parasitemia, the estimated parasitemia generated by their entire suggested method was not particularly close to the truth. In every study so far, a straightforward segmentation technique was paired with a classifier built on the CNN. With CNN, erythrocyte segmentation is also feasible.

Using a network design where convolutional layers are followed by deconvolutional layers to build a segmentation mask for slide images, Delgado-Ortet *et al*. (2020) applied this to the classification of thin smear images. They used this with an eight-layer CNN to classify the segmentation output, resulting in a segmentation accuracy more than the test set of 93.72% and a specificity for malaria identification of 87.04%. The Caffe CNN architecture, which employs 3 fully linked layers come after 5 convolutional layers is used as a feature extractor in Mehanian *et al.* (2017), a technique for the classification of Giemsa-stained thick blood smear pictures. After being fed into a logistic regression classifier, the network's candidate objects are divided into parasites and non-parasites with a sensitivity of 91.6% and a specificity of 94.1%.

Table 2.1 Summary of previous results

ML Method	Segmentation Method	Performance	Classification Method	Performance	References
Supervised Learning	Otsu thresholding + watershed	Accuracy 99.43%	SVM	Sensitivity=93.12 % Specificity=93.17 %	Savkare et al (2018).
Supervised Learning	Marker controlled watershed	-	SVM with feature selection, classification of species/stage	Accuracy=84%	Das. et al. (2013)
Supervised Learning	Threshold + Hough circle transform	-	CNN: GoogLeNet	Accuracy= 98.17%	Dong et al. (2017)
Supervised Learning	Level-set based algorithm	-	Custom CNN	Sensitivity=93.12 % Specificity=95.12 %	Rajamaran et al. (2018)

Supervised Learning	Two stage thresholds	-	CNN	Sensitivity=96.98 % Specificity=98.50 %	Gopakumar et al. (2018)
Supervised Learning	Convolutional + Deconvolutional NN	Accuracy 93.72%	Custom CNN	Specificity=87.04 %	Delgado-Ortet et al. (2020)

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1 Materials

The equipment and software utilized during this work are listed below:

1 A laptop (64-bit operating system, 4.00 GB RAM, 1.10GHz)

2 Orange Software (Version 3.23.0)

3.1.1 Malaria datasets

The performance of the comparison models was assessed using the NIH Malaria dataset, which is publicly available on the National Library of Medicine, National Institutes of Health (NIH). Lister Hill National Center for Biomedical Communications site at https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html. There are 27,558 cell pictures in the dataset, 13,779 of which were parasitised and 13,779 of which are uninfected. Plasmodium-containing cells are referred to as parasitized, whilst healthy cells devoid of Plasmodium are referred to as uninfected. The colour distributions of the cell pictures vary as a result of various bloodstains that appeared during the data collection process. From the malaria dataset, examples of segmented images of infected and uninfected red blood cells are displayed in Figures 3.4 and 3.5.

3.2 Methods

3.2.1 Segmentation method

The Giemsa-stained thin films images were embedded using Artificial Neural Network based segmentation method: SqueezeNet, a deep model for image recognition that uses 50 times less parameters than AlexNet to attain accuracy on ImageNet, it automatically retrieves image

vectors from the server, the image embedding widget was connected to the import images widget.

3.2.2 Classification methods

In this study, four supervised learning classifiers were used, Logistic Regression, Random Forest, SVM and KNN and two unsupervised learning classifiers were used namely hierarchical clustering and K-means. They are succinctly explained as follows:

Random Forest Algorithm: During the classification process, it is aimed to increase the classification value by using more than one decision tree. Instead of producing a single decision tree, Breiman (2001) suggested integrating the judgments instead of building a single decision tree, of a huge number of multivariate trees, each trained with a separate training set. Training sets are produced from the initial training set using bootstrapping and random feature selection. The class that receives the most votes in the decision forest was taken to be the final conclusion, and it includes the entering test data. Each choice tree then offers its own conclusion.

K-Nearest Neighbors Classifier (KNN): A fresh sample is classified using this supervised learning technique based on the adjacent training samples that were already present in the feature field. The test data is mapped to the class that the k neighbors share most when it is given (Acharya *et al.*, 2012).

Support Vector Machines (SVM): This supervised classification method produces a separating hyperplane in high dimensional space that was used for classification. For any class, the hyperplane that was farthest from the nearest training data point gets a respectable separation (Acharya *et al.*, 2012).

Logistic regression: is a technique for calculating the likelihood of a discrete output given an input variable. One with a binary result, such as true or false or yes or no, is the most common type of logistic regression model. (Edgar and Manz, 2017).

k-means: is employed to divide the cases or variables in a dataset into non-overlapping groups, or clusters, in accordance with the traits discovered. It is preferred that groupings of cases or variables have a high degree of similarity within each group and a low degree of similarity between them. (Friedman *et al.,* 2001).

Hierarchical clustering: creates a tree over the data, frequently in binary. The leaves are individual data objects, while the root is a single cluster that contains all of the data. Between the root and the leaves are intermediary clusters that contain subsets of the data. The fundamental purpose of hierarchical clustering is to produce "clusters of clusters" that advance higher in order to construct a tree.

3.3 Supervised learning

The methodology flow chart for supervised learning used for this study is presented in figure 3.1. This presents an outline of the steps that will be followed to address the problem to be solved in this study.

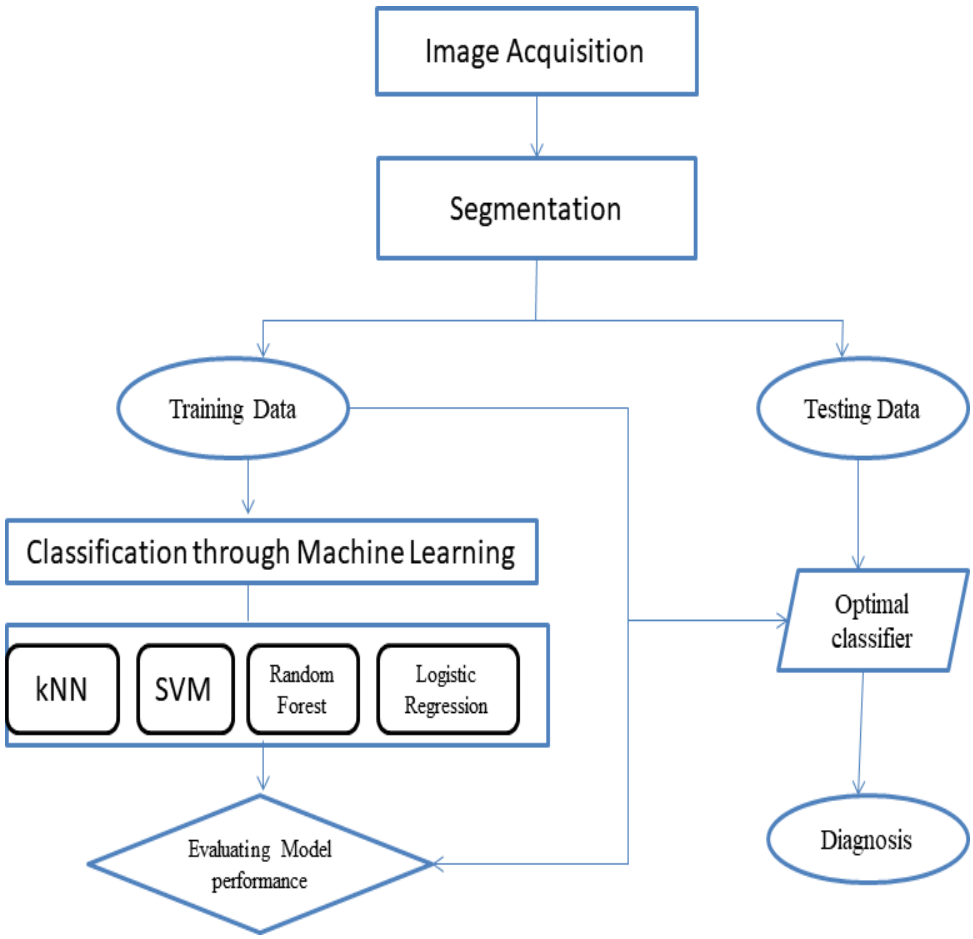


Figure 3.1: Methodology flow chart for supervised learning

The specific steps followed in implementing supervised machine learning are outline as follows.

STEP I:

1. Orange Software was downloaded from Google.

STEP II:

1. The Orange Software was launched on the system by clicking on the Orange application downloaded from Google

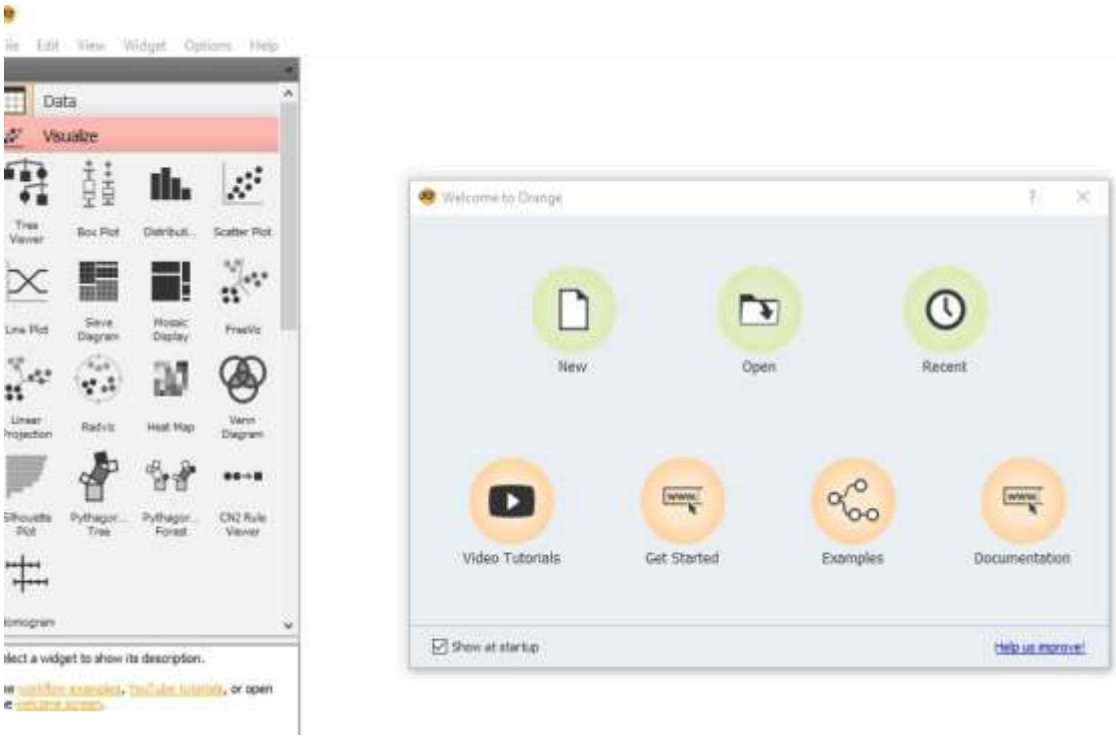


Figure 3.2: Orange canvas.

STEP III:

This stage describes how the images were imported

1. To import images, the import image widget on the orange canvas was clicked on.
2. The import image widget was double clicked to import the labelled images for the supervised learning.
3. The image viewer was connected to the import image to check the content of the directory.

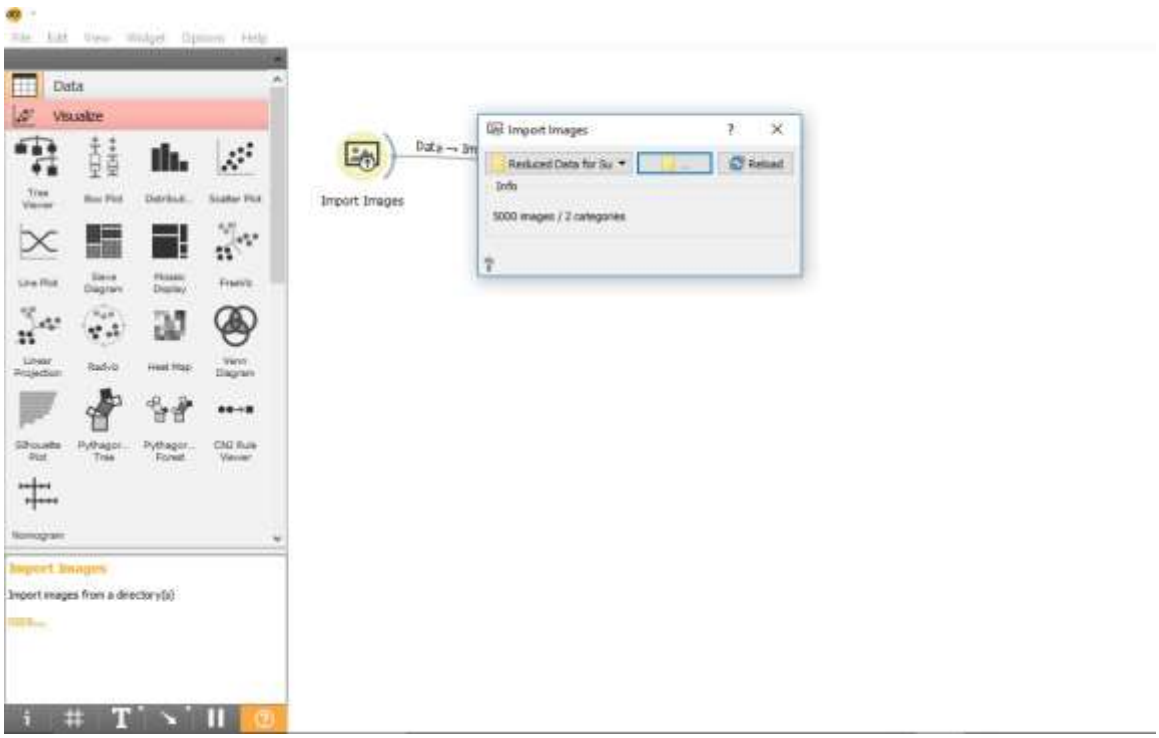


Figure 3.3: Import image using the import images widget

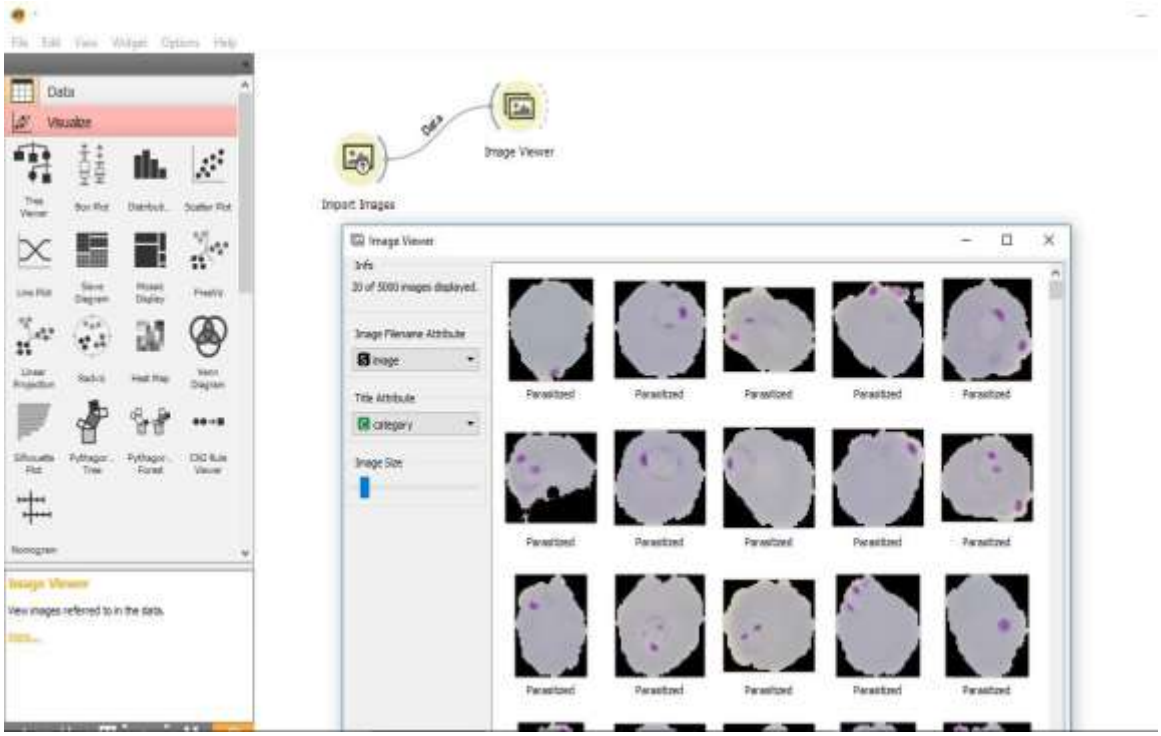


Figure 3.4: Samples of parasitised images.

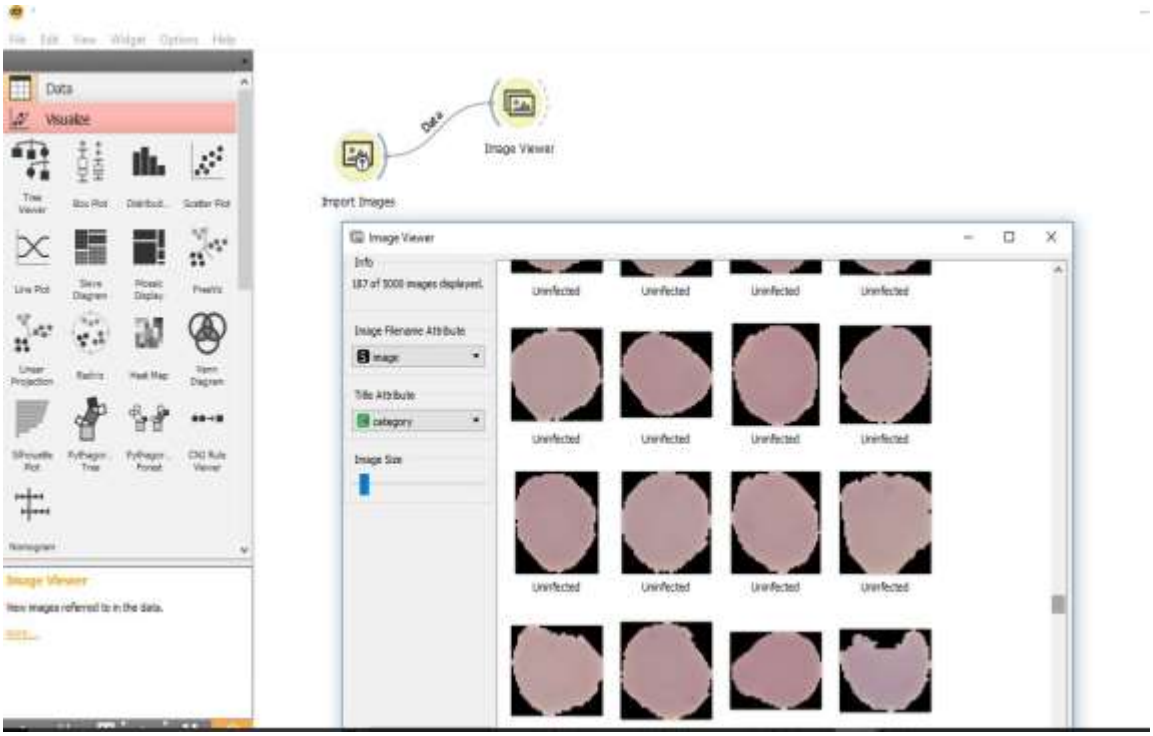


Figure 3.5: Sample of uninfected images.

STEP IV:

The image embedding widget is the most important for the image analytics; Classification and regressions tasks requires data in the form of numbers. Image Embedding widget works by converting images to vectors of numbers.

1. Import images was connected to the image embedding widget for the server to push image through a pre-trained deep neural network and return number vectors to the widget.
2. Image Embedding widget was connected to Data Table widget to see vector representation of the images.

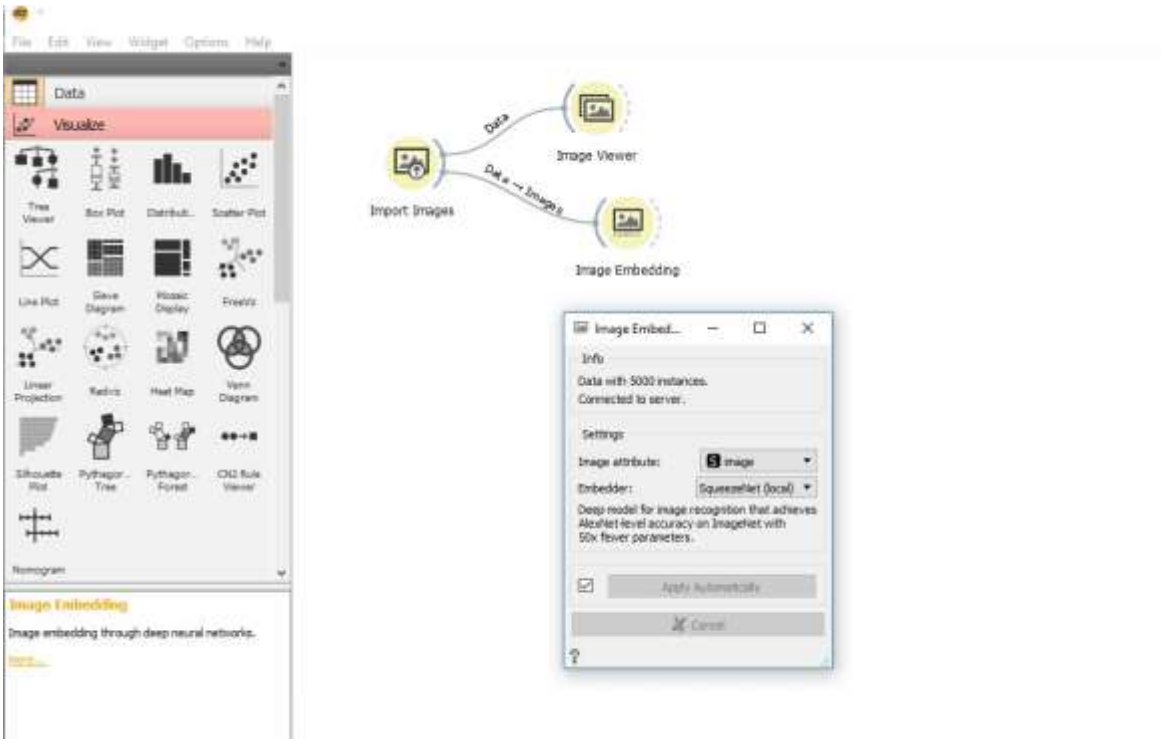


Figure 3.6: Image embedding.

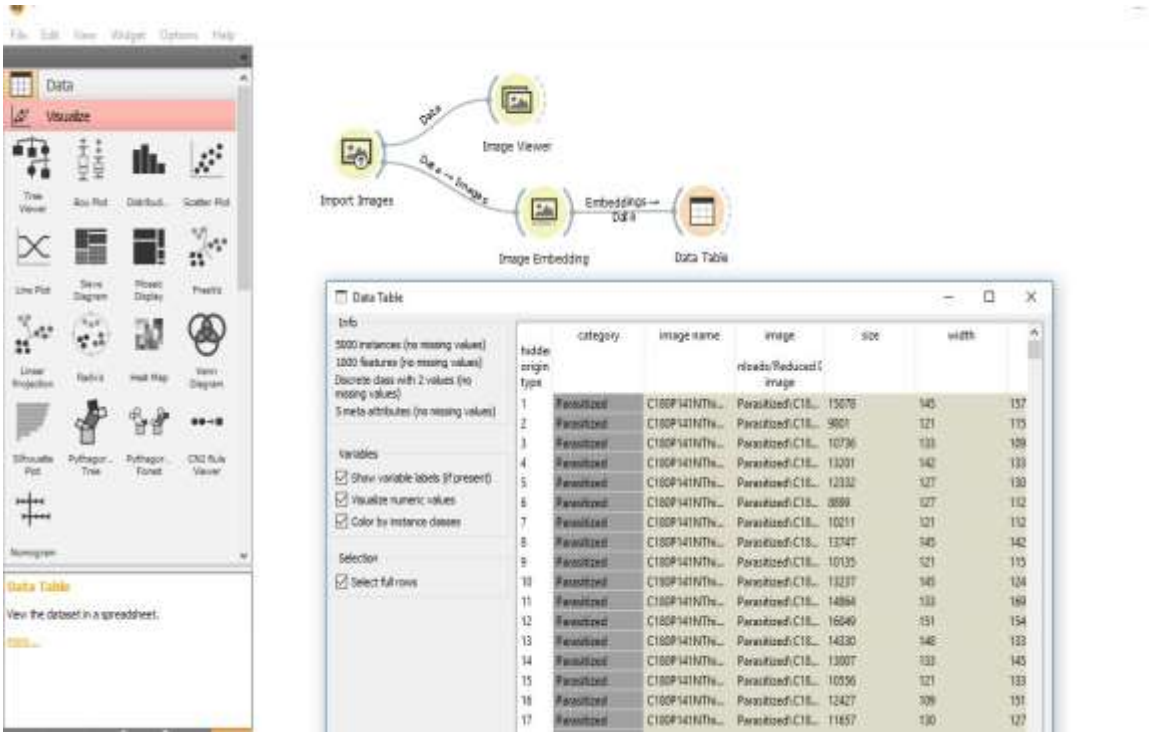


Figure 3.7: Data table.

STEP V:

Sampling the data to split data into test and training data sets

1. The Image Embedding widget was connected to the Data Sampler widget
2. The data was divided into equal segments keeping 70% of the data instances in the sample.
3. Sample Data was clicked on to process the output.
4. Two outputs of Data Table were connected to Data Sampler: Data Sample -> Data and Remaining Data -> Test Data.
5. The Data Table was renamed as Test Data and Train Data.

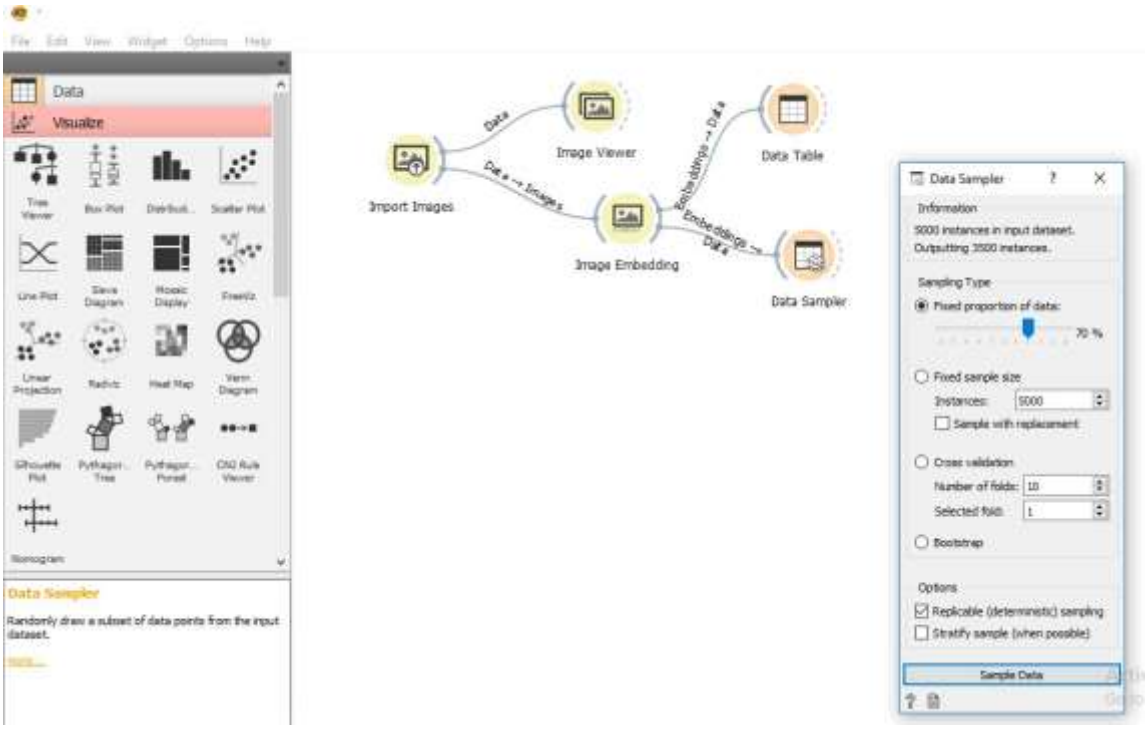


Figure 3.8: Data sampler.

The complete amount of data was split into training and test data for the purpose of training. The remaining 30% of the data 8267 images is included in the test data, leaving the training

data with 70% of the total data of 19291 images.

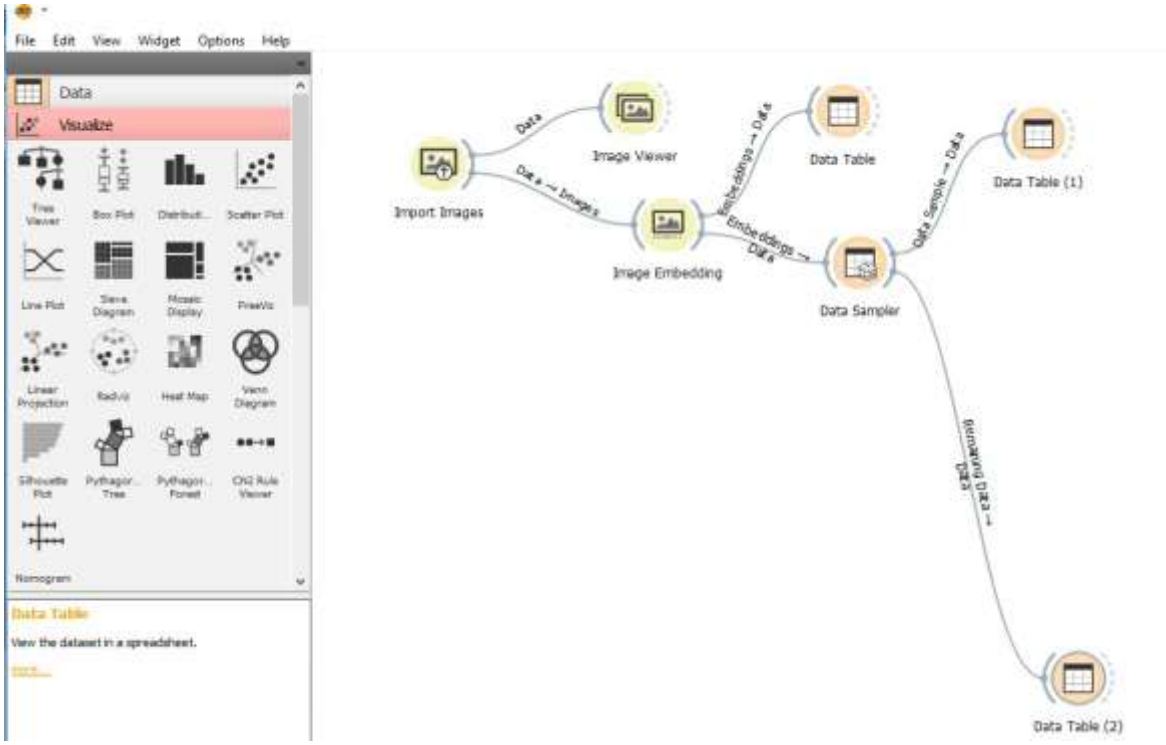


Figure 3.9: Divided into train and test data by the data sampler.

Figure 3.10: Train data.

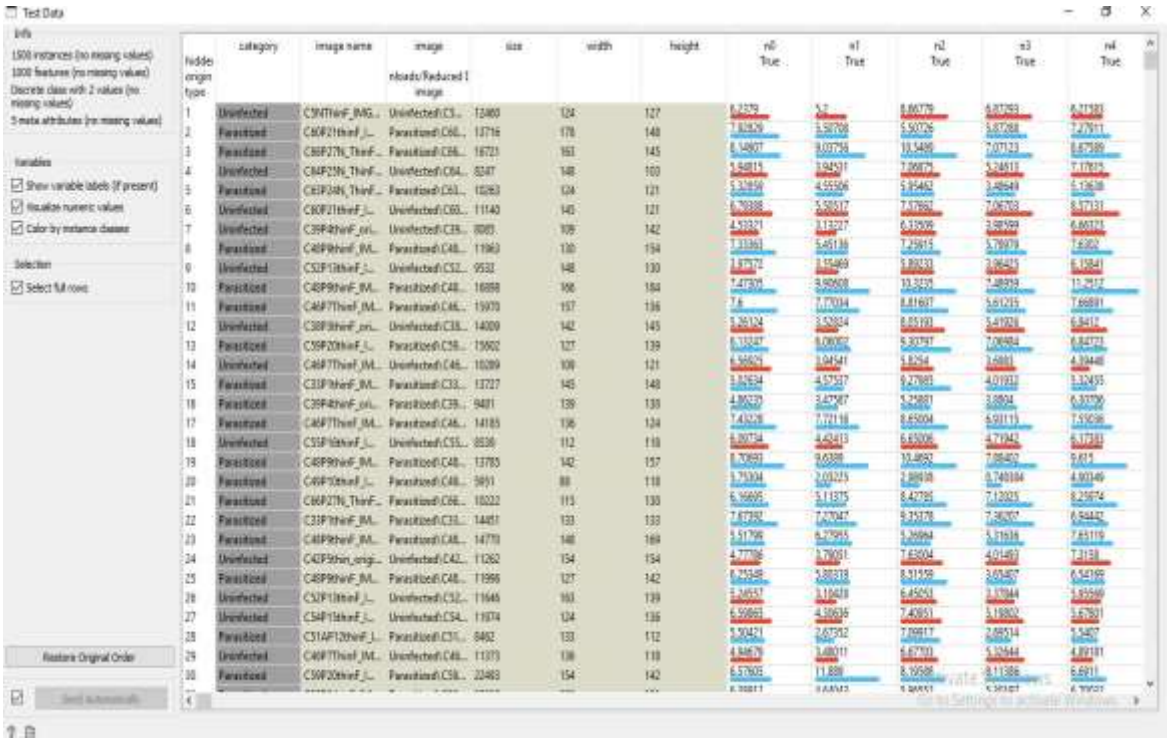


Figure 3.11: Test data.

STEP VI:

The learners responsible for classification and regression were introduced.

1. The train data widget was connected to Test and Score widget.
2. Data Sampler was connected to Models.
3. Logistic Regression was connected to Data Sampler widget and output to Test and Score widget.
4. K-Nearest Neighbours (KNN) was connected to Data Sampler widget and output to Test and Score widget.
5. Support Vector Machine (SVM) was connected to Data Sampler widget and output to Test and Score widget.
6. Random Forest was connected to Data Sampler widget and output to Test and Score widget.

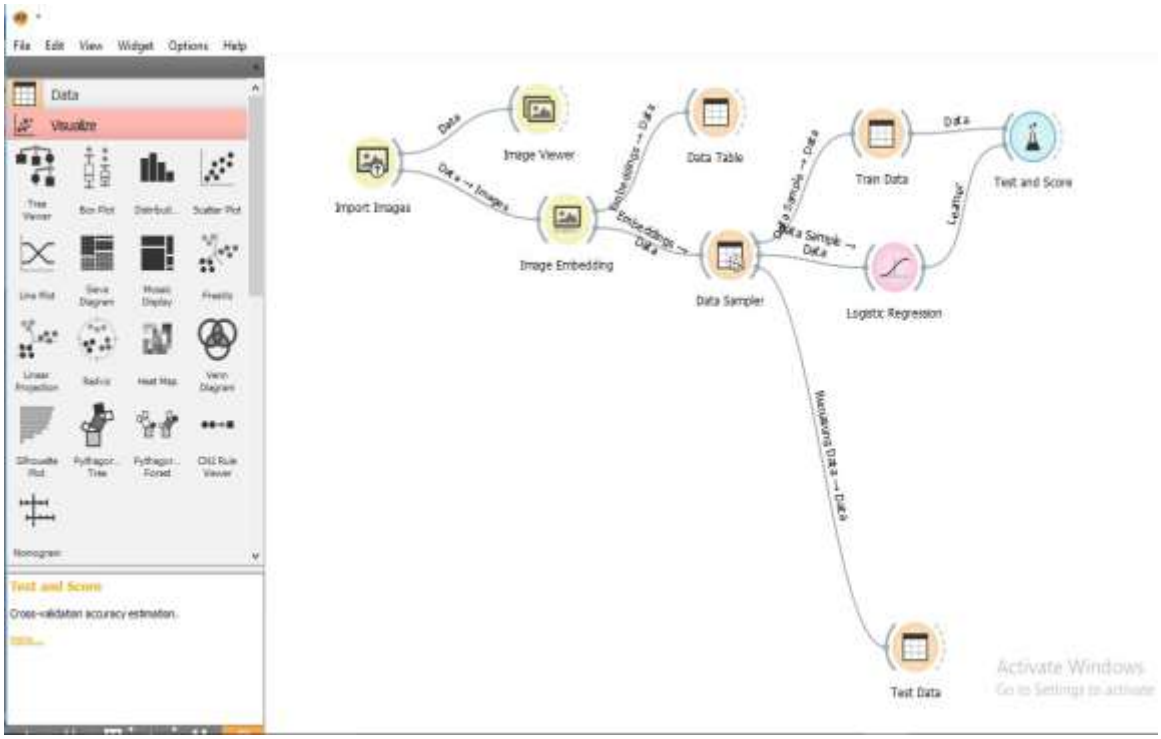


Figure 3.12: Connecting Logistic regression as classifier

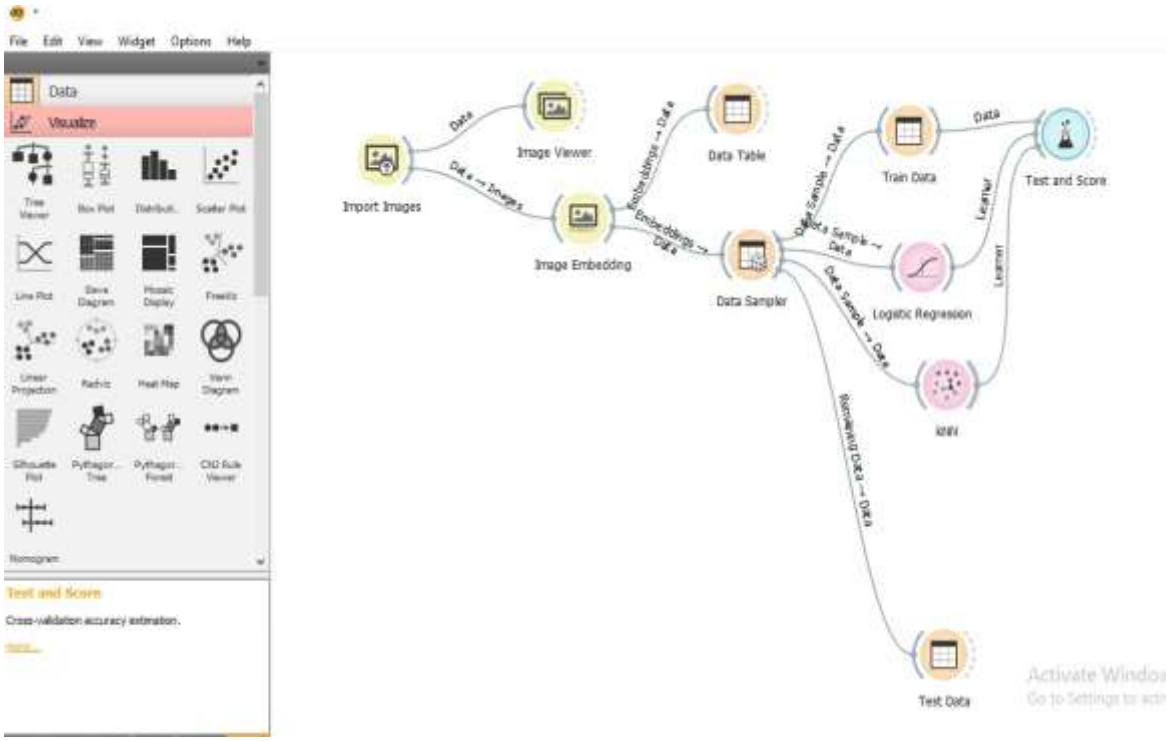


Figure 3.13: Connecting kNN as classifier.

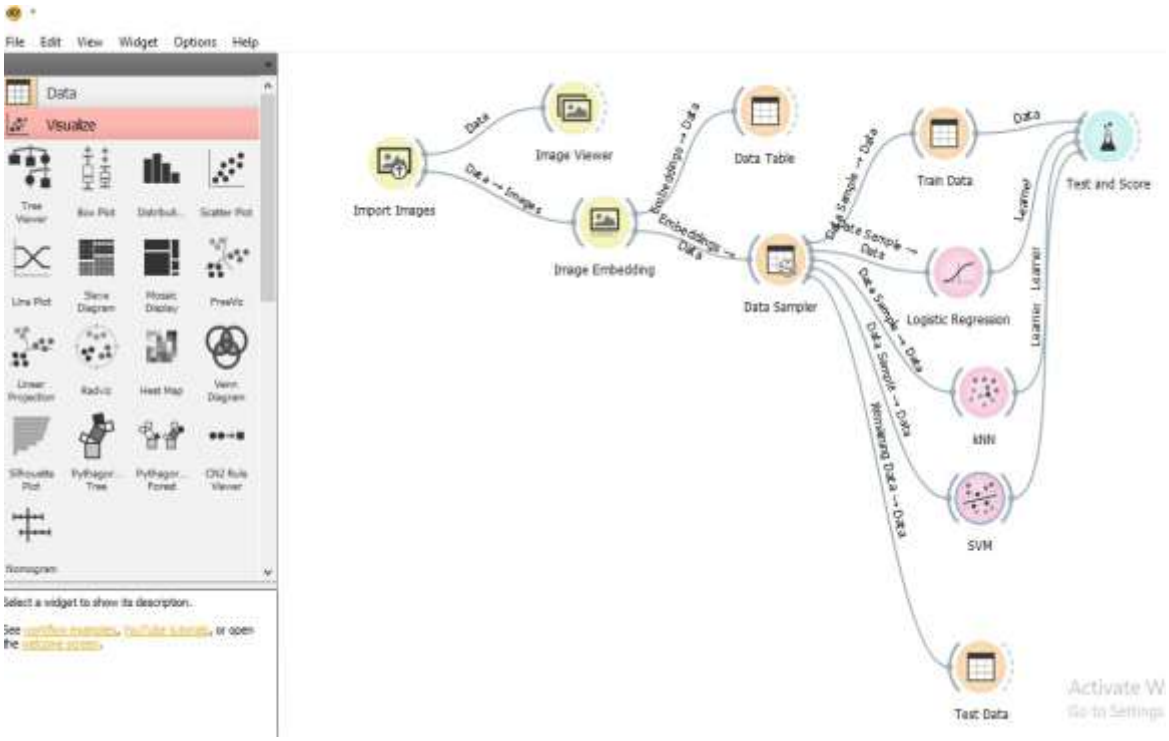


Figure 3.14: Connecting SVM as classifier.

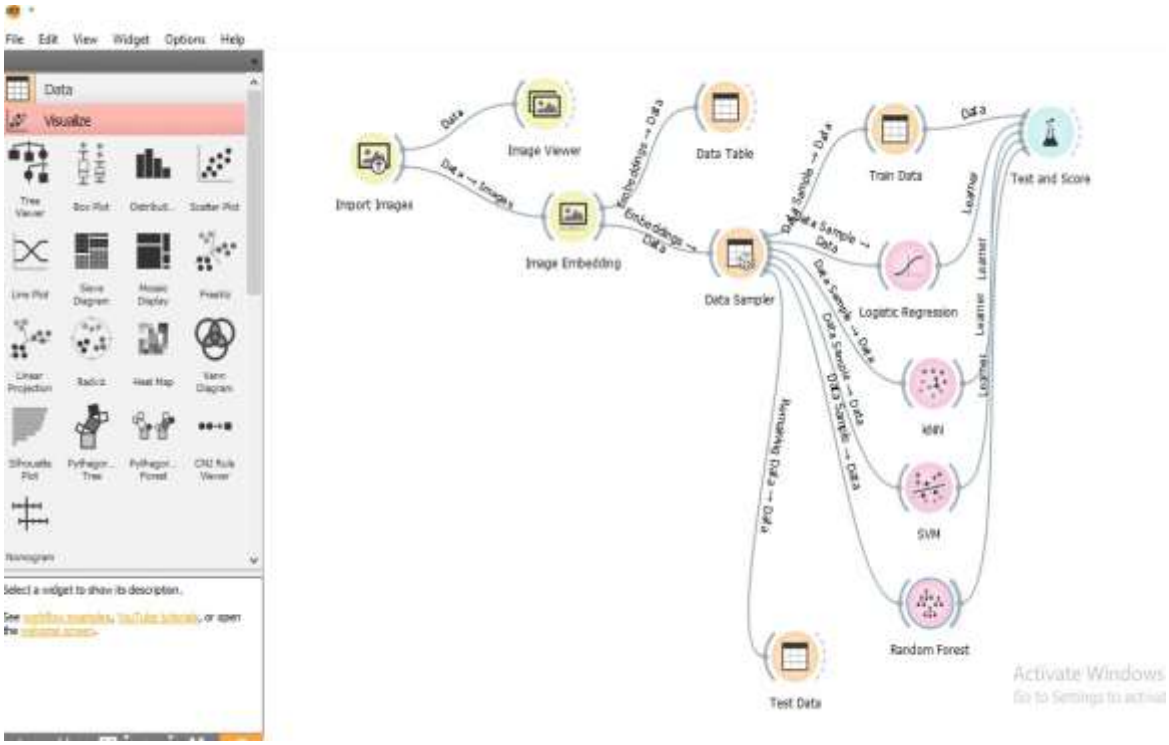


Figure 3.15: Connecting Random Forest as classifier.

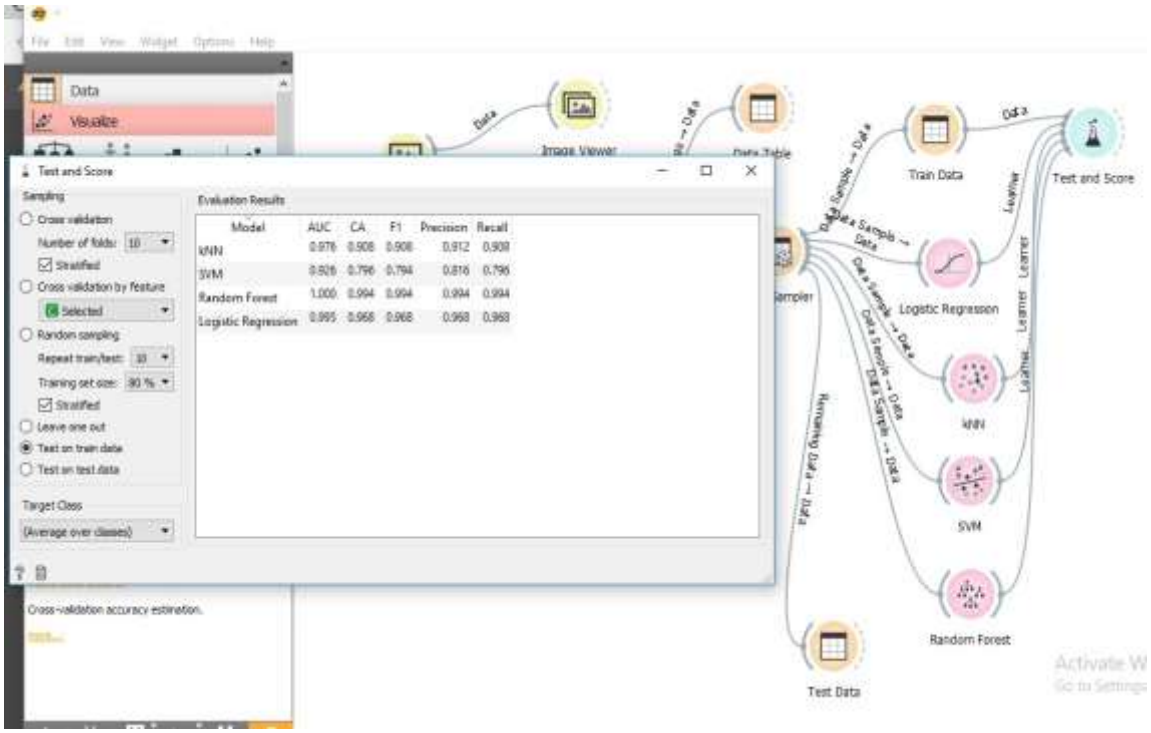


Figure 3.16: Evaluation results from test and score.

STEP VII:

To obtain prediction using the Test Data

1. Test Data was connected to the Prediction widget.
2. Logistic Regression was connected to the Prediction widget.
3. Random Forest was connected to the Prediction widget.

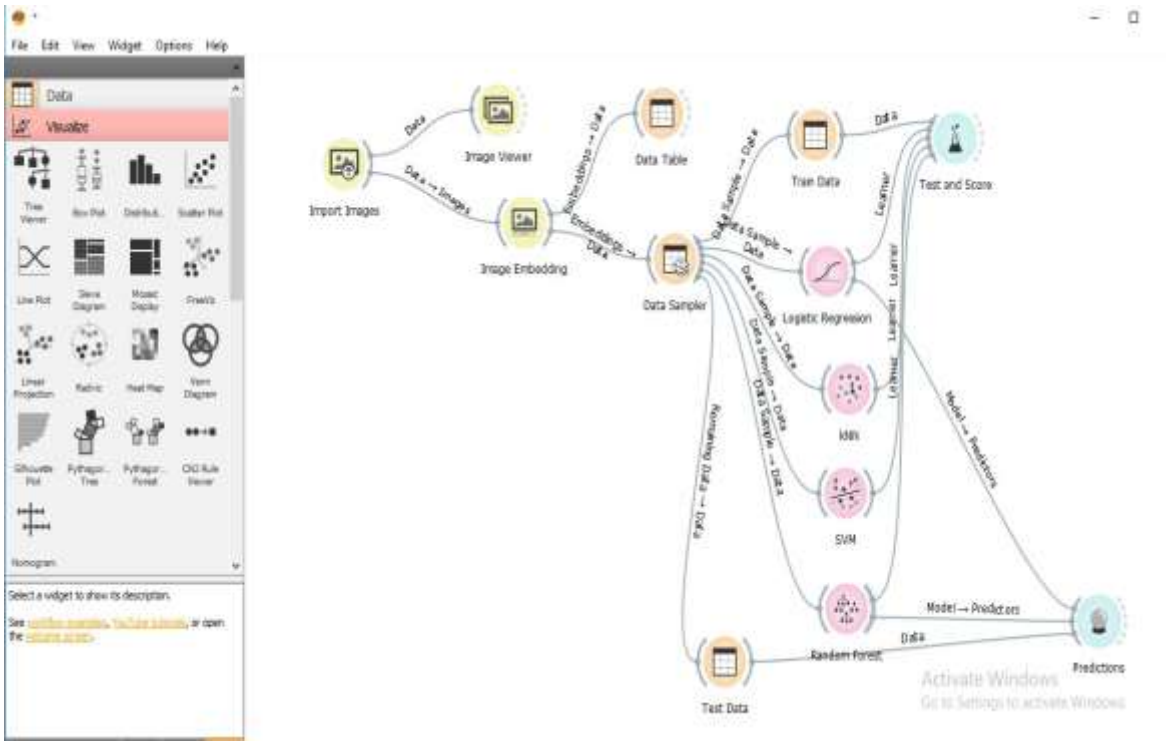


Figure 3.17: Prediction connected to test data and classifiers.

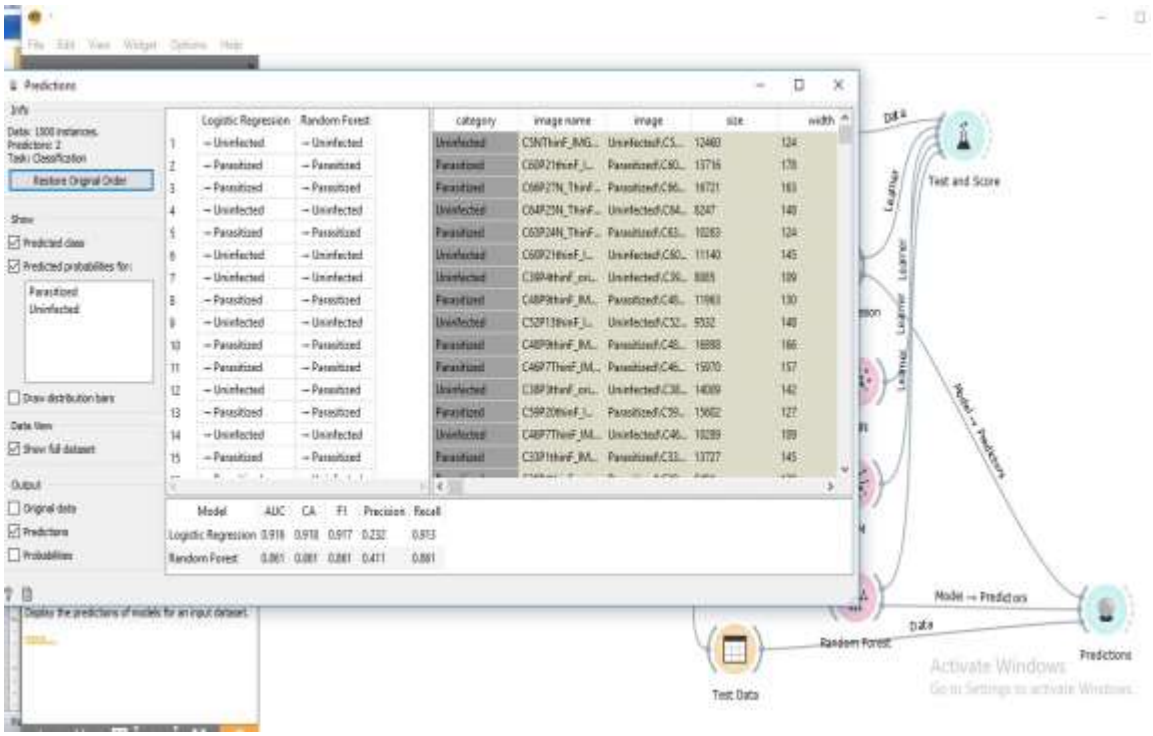


Figure 3.18: Prediction result.

3.3.1 Model evaluation metrics

The Test and Score performs a tenfold cross validation on images and reports on accuracy, It assesses the accuracy of each model by comparing the results of the target variables to the actual data. The Area Under Receiver-Operator Curve (AUC), Classification Accuracy (CA), F1, Precision, and Recall measures a model's performance.

AUC: AUC's value ranges from 0 to 1. The AUC of a model with 100% erroneous predictions is 0.0, while the AUC of a model with 100% correct predictions is 1.0.

Accuracy: is the percentage of accurate predictions made by the model. The official definition of accuracy is as given in equation 3.1:

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction} \quad (3.1)$$

For binary classification, accuracy can also be assessed in terms of positives and negatives, as shown below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.2)$$

Precision: efforts to determine what percentage of affirmative identifications were actually accurate. Following is a definition of precision:

$$Precision = \frac{TP}{TP+FP} \quad (3.3)$$

Recall tries to address the following query: What percentage of real positives were successfully identified? Recall is defined mathematically as follows:

$$Recall = \frac{TP}{TP+FN} \quad (3.4)$$

F1 score demonstrates how precision and recall are balanced. F1 score is described mathematically as shown in Equation 3.5:

$$F1\ score = 2 * \frac{(precision*recall)}{(precision+recall)} \quad (3.5)$$

Sensitivity is a statistic used to assess how well a model can forecast true positives for each accessible category as shown in Equation 3.6.

$$sensitivity = \frac{TP}{TP+FN} \quad (3.6)$$

Specificity is the statistic used to determine how well a model predicts true positives for each important category as shown in Equation 3.7.

$$specificity = \frac{TN}{TN+FP} \quad (3.7)$$

where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

3.4 Unsupervised Learning

The methodology flowchart for unsupervised machine learning to be implemented in this study is presented in Figure 3.19

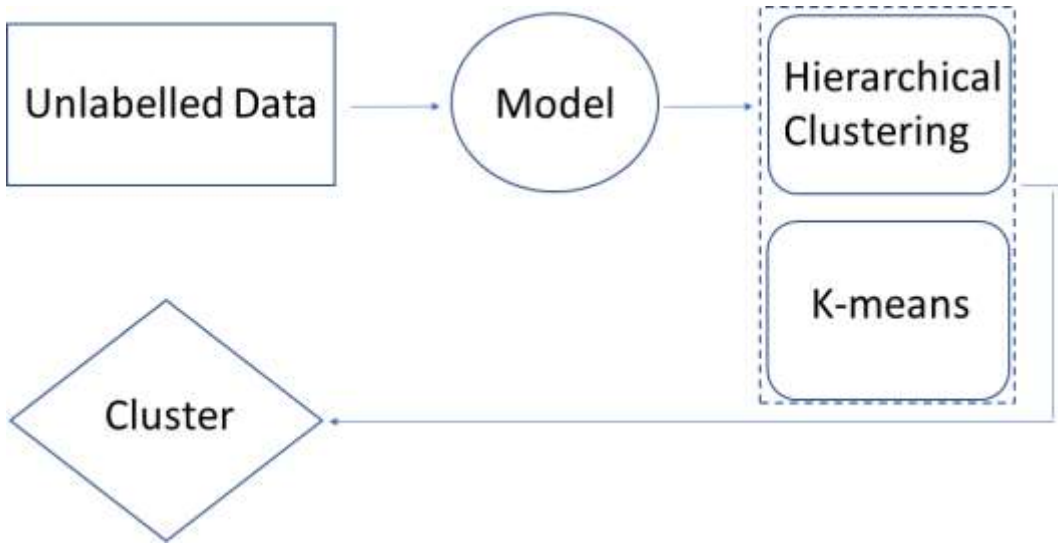


Figure 3.19: Methodology flowchart for unsupervised learning.

The following specific steps were employed in performing unsupervised machine learning on the images.

STEP I:

1. Unlabelled data were used.

STEP II:

1. The Orange Software was launched on the system

STEP II:

This stage is where the image is imported

1. The import image widget was clicked on the orange canvas to import images.

2. the unlabelled images were imported using the import image widget.
3. The image viewer was connected to the import image to check the content of the directory.

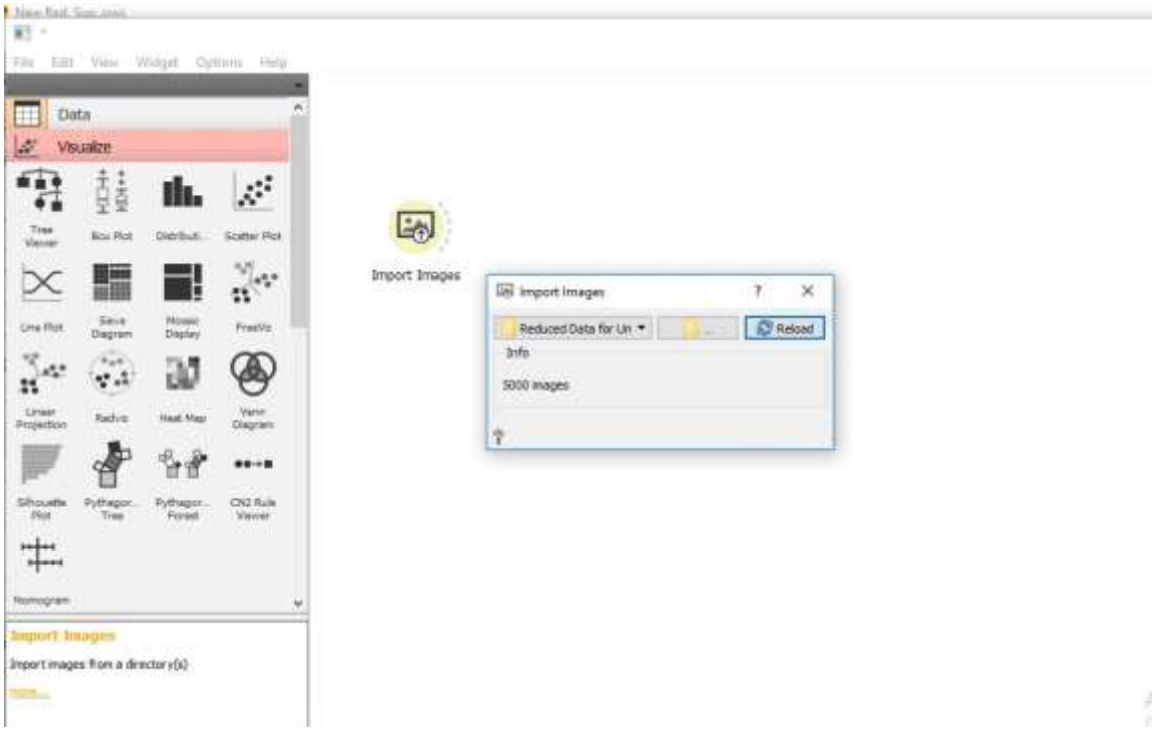


Figure 3.20: Import unlabelled images.

STEP IV:

i. step IV of section 3.2.1 was repeated.

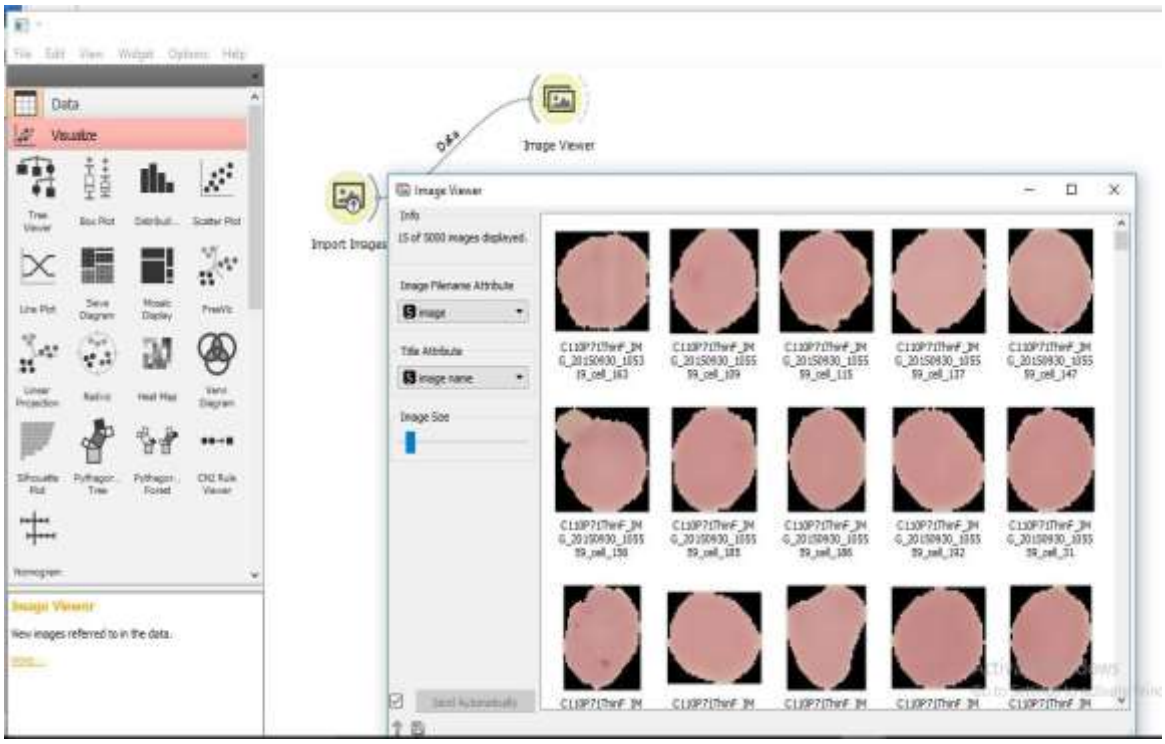


Figure 3.21: Unlabelled images.

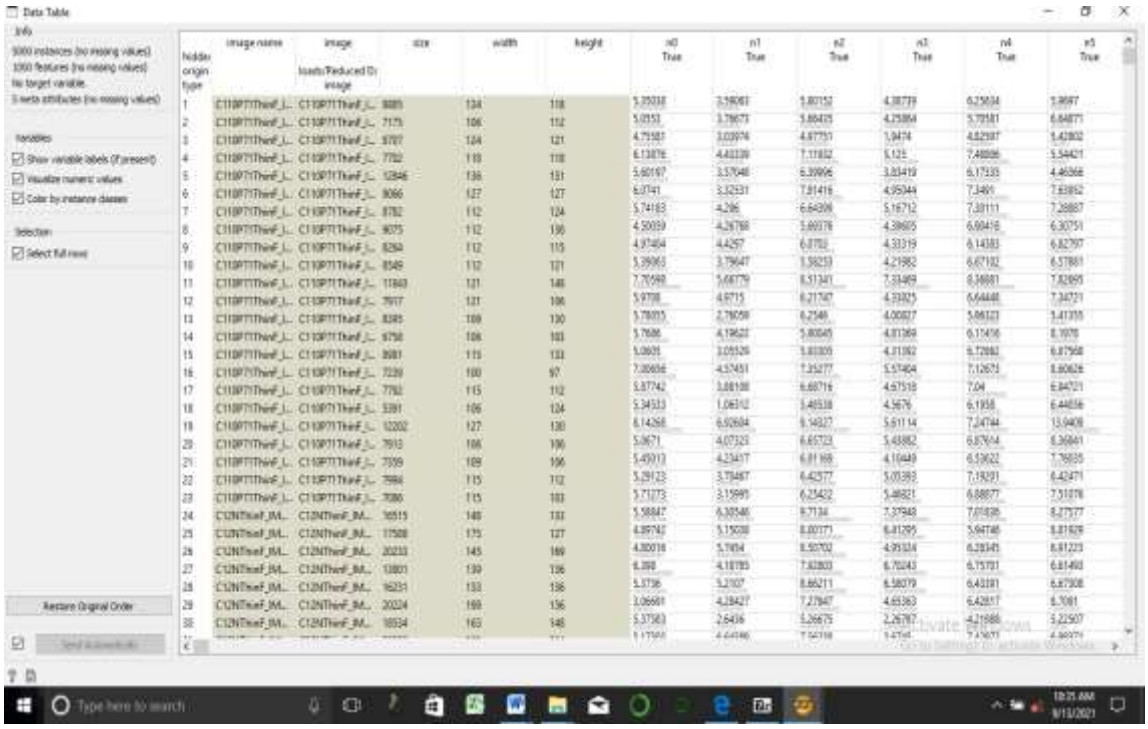


Figure 3.22: Data table

STEP V:

To find clusters using the method of k-means

1. The Image Embedding widget was connected to K-means

STEP VI:

To discover groups or subgroups using Hierarchical Clustering

1. Distances widget was connected to Image Embedding
2. Hierarchical Clustering widget was connected to Distances widget.

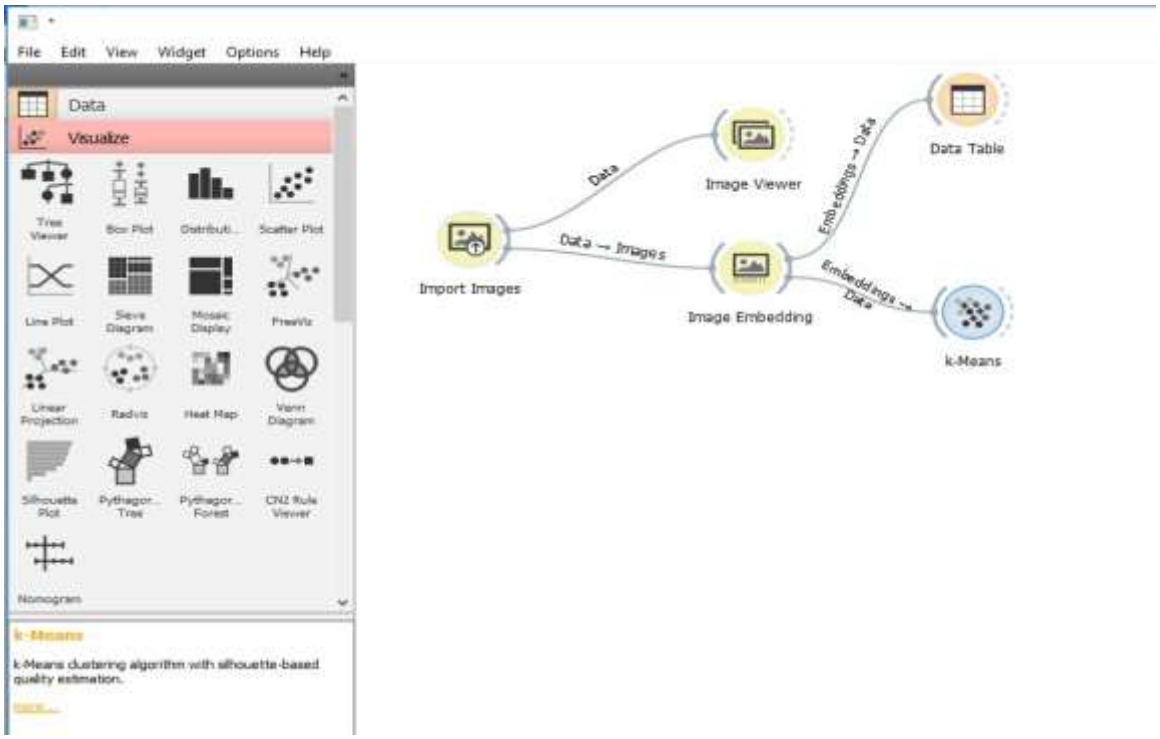


Figure 3.23: k-means implementation.

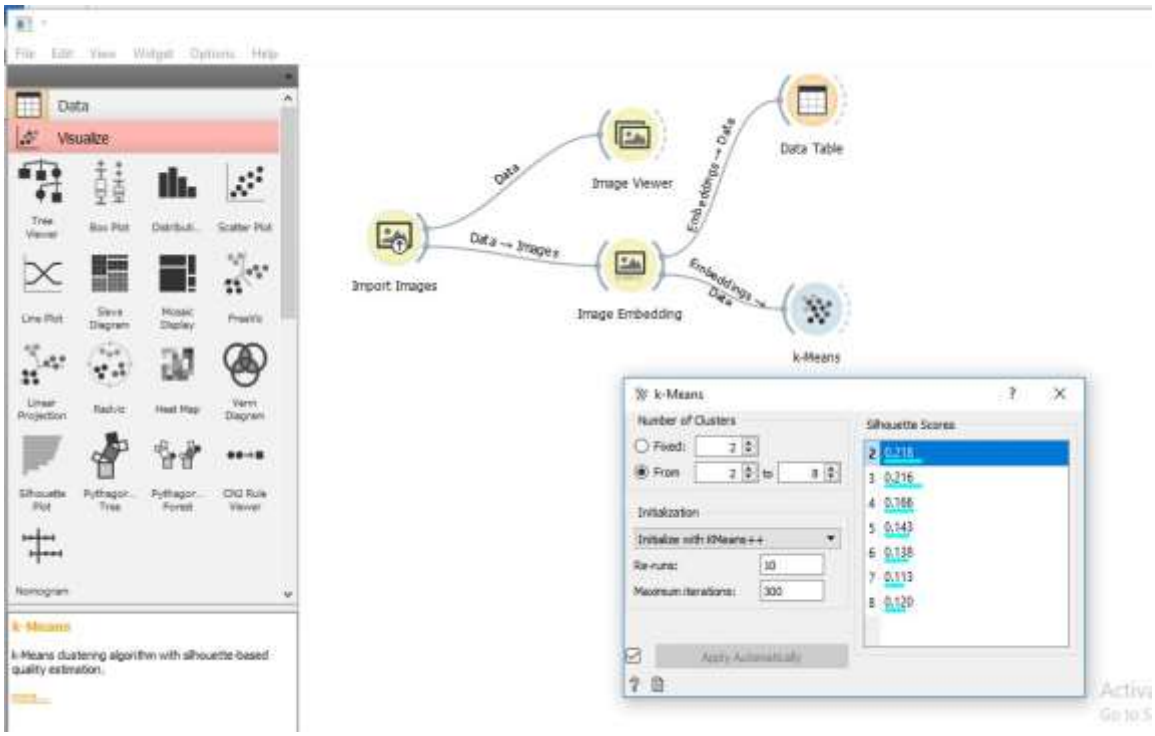


Figure 3.24: Silhouette scores.

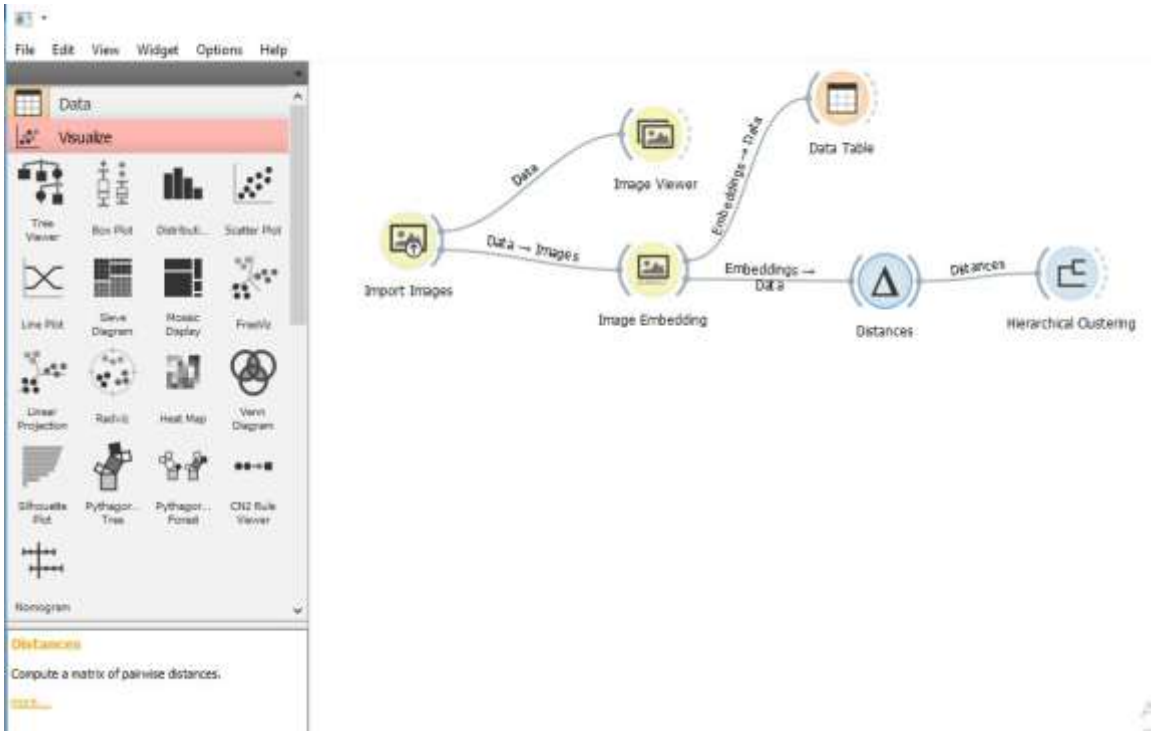


Figure 3.25: Hierarchical clustering implementation.

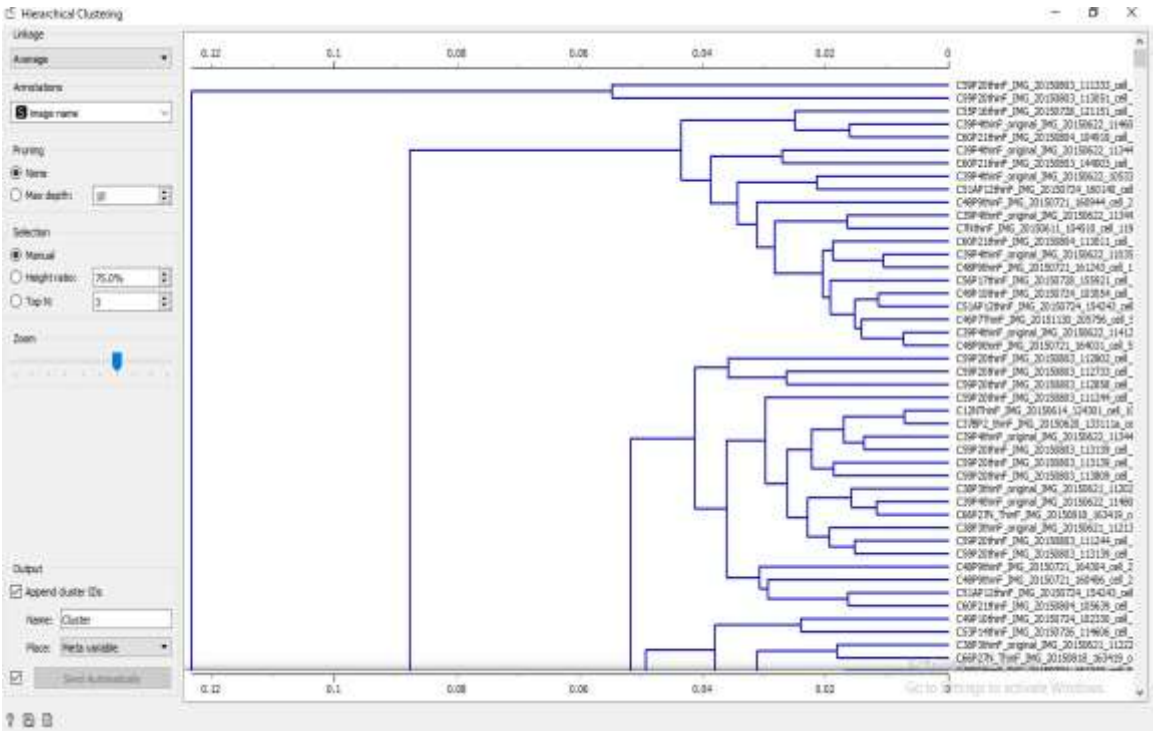


Figure 3.26: Dendogram implementation.

CHAPTER FOUR

4.0 RESULTS AND DISCUSSION

4.1 Supervised Learning

Four alternative methods were looked into in order to identify the best and most effective model. The models include KNN, SVM, Random Forest, and Logistic Regression.

Considering the percentage of correctly categorized target variables in our model's classification accuracy. Table 4.1 shows that the accuracy of the KNN is 92%, that of SVM is 78%, that of the Random Forest is 99%, and that of Logistic Regression is 95%. The most accurate machine learning models for this dataset are produced by Random Forest techniques (99%) followed by Logistic Regression (95%).

Table 4.1: Evaluation Results for different model used (KNN, SVM, Random Forest and Logistic Regression) for 27558 images

Model	AUC	CA	F1	Precision	Recall
kNN	0.985	0.921	0.920	0.925	0.921
SVM	0.918	0.788	0.782	0.821	0.788
Random forest	1.00	0.993	0.993	0.993	0.993
Logistic Regression	0.990	0.955	0.955	0.955	0.955

To obtain the prediction with the remaining 30% test data, we connect our classifiers that are the logistic regression and the random forest to the prediction widget.

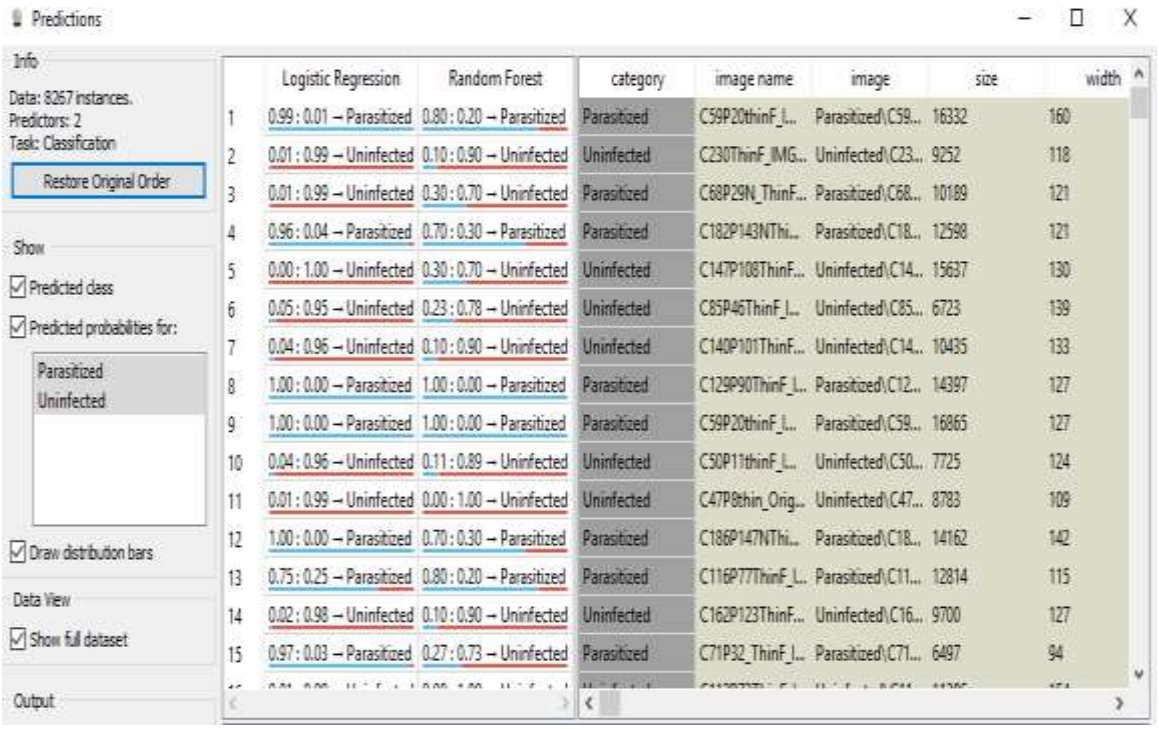
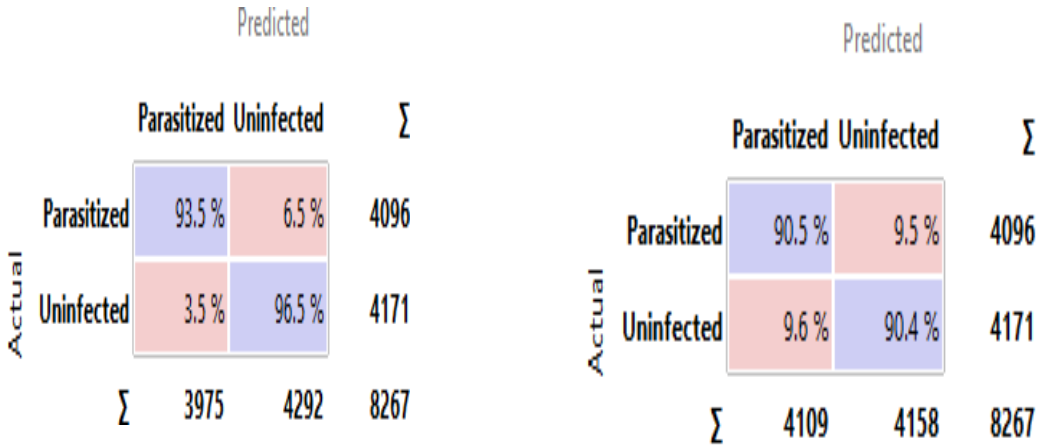


Figure 4.1: Prediction results for the test data using 30% of the remaining data from 27558 images.

Figure 4.1 shows the prediction results of the test data using 30% of the remaining data, the data consist of 8267 instances, 2 predictors (Logistic Regression and Random Forest), from the prediction result it was observed that Logistic Regression and Random Forest classifier were able to predict correctly most parasitised and uninfected images, however there are few cases in which misclassification occurred. To get the proportion of instances between the predicted and actual class, the confusion matrix was introduced.

Confusion matrix reports on actual image classes and predicted classes and provides a data instance count for each combination. True Positive and True Negative is highlighted with blue while the misclassified which are False Positive and False Negative are reported with pink.

From the confusion matrix in Figure 4.2a, while using logistic regression as the classifier, it shows that 3.5% of the data was actually uninfected but was predicted as parasitised (False Positive) while 6.5% of the data that was actually parasitised was labelled as uninfected (False Negative). However, 93.5% of parasitised images were correctly predicted (True Positive) and 96.5% of the uninfected images were correctly predicted (True Negative). While from Figure 4.2b using random forest as the classifier, it shows that 9.6% of the data was actually uninfected but was predicted as parasitised (False Positive) while 9.5% of the data that was actually parasitised was labelled as uninfected (False Negative). However, 90.5% of parasitised images were correctly predicted (True Positive) and 90.4% of the uninfected images were correctly predicted (True Negative).



(a) (b)

Figure 4.2: Predicted and actual parasitised red blood cells as shown in Confusion matrix of 27558 images using (a) logistic regression (b) Random Forest as classifier.

For Logistic Regression:

$$sensitivity = \frac{TP}{TP+FN} = \frac{93.5}{93.5+3.5} = \frac{93.5}{97} = 96.4\% \quad (4.1)$$

$$specificity = \frac{TN}{TN+FP} = \frac{96.5}{96.5+6.5} = \frac{96.5}{103} = 94\% \quad (4.2)$$

For Random Forest:

$$sensitivity = \frac{TP}{TP+FN} = \frac{90.5}{90.5+9.6} = \frac{90.5}{100.1} = 90.4\% \quad (4.3)$$

$$specificity = \frac{TN}{TN+FP} = \frac{90.4}{90.4+9.5} = \frac{90.4}{99.9} = 90.5\% \quad (4.4)$$

The logistic regression was observed to have reached sensitivity and specificity. of 96.4% and 94% respectively in predicting parasitised cells and uninfected cells and significantly outperformed the random forest classifier that obtained sensitivity of 90.4% and specificity of 90.5%. Comparing with methods that were described in Table 2.1, this method outperforms the one that was proposed by Rajaraman *et al.* (2018) (sensitivity 93.12%), and the one proposed by Savkare *et al.* (2016) (sensitivity = 93.12% and specificity = 93.17%), and the one proposed by Delgado *et al.* (2020) (specificity 87.04%)

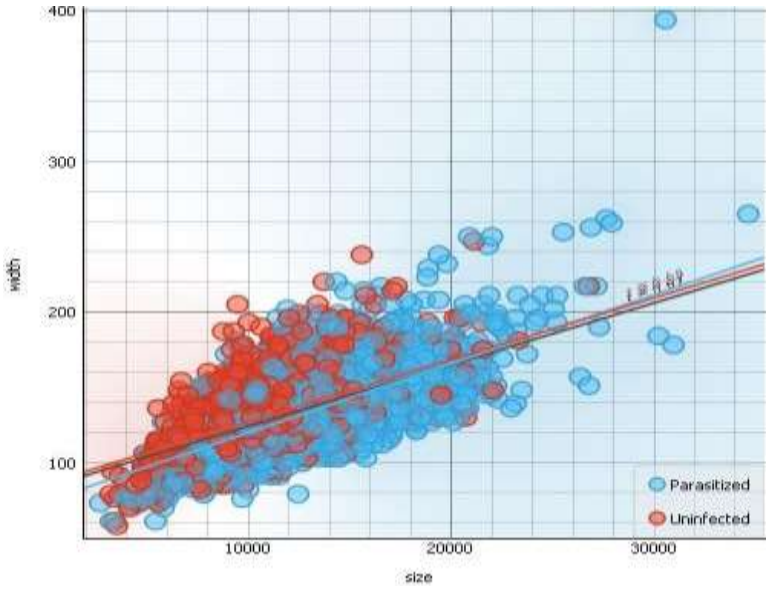
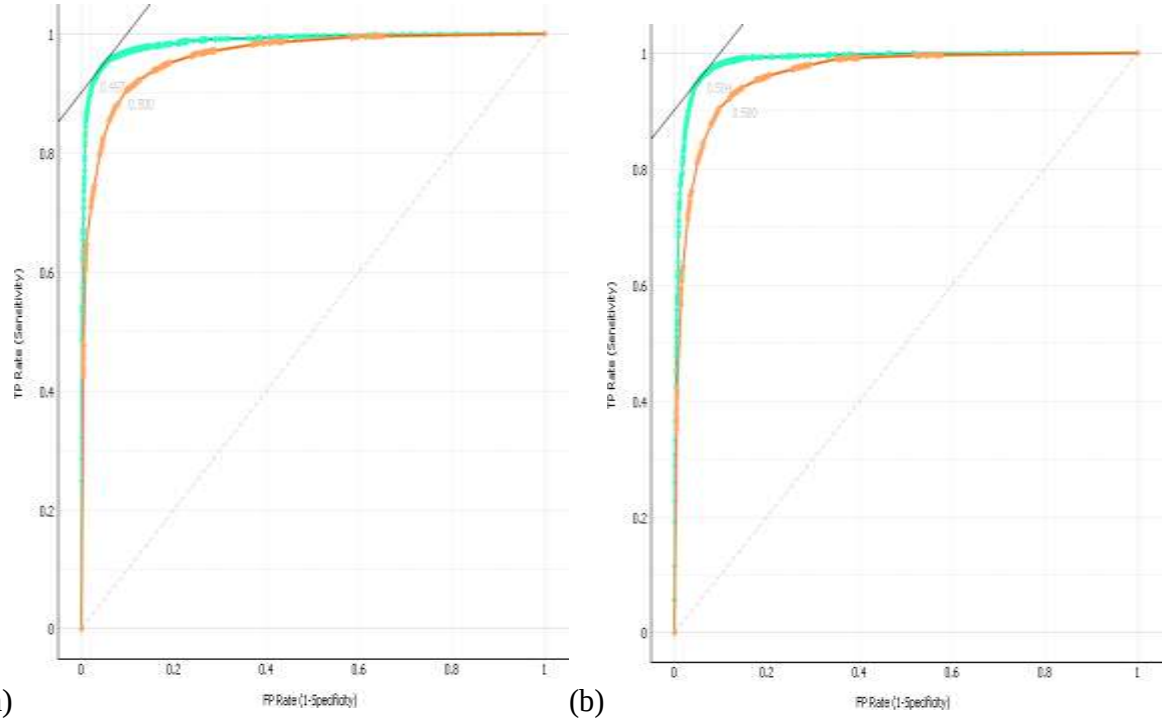


Figure 4.3: Scatter plot displaying a 2 dimensional scatter plot visualisation of the whole dataset.

Figure 4.3 displays a 2 dimensional scatter plot visualisation of the 27558 thin smear Geimsa-stained images, The data is represented as a set of points, where each point's size value on the

x-axis determines its location on the horizontal axis and its width value on the y-axis determines its position on the vertical axis. The data points for parasitised images are represented in blue and data points for uninfected images are represented in red.



(a) (b)

Figure 4.4: True positive rate against false positive rate shown in the ROC analysis for (a) parasitised (b) and uninfected photos (whole dataset).

On the x-axis of Figure 4.4 (a,b) curve are graphs of the false positive rate (1-specificity; likelihood that the target is 1 when the true value is 0). Sensitivity (probability that the goal equals 1 when the true value equals 0) is plotted against the true positive rate on the y-axis. The proximity of the curve to the top and left borders shows how precise the classifiers are. The ROC analysis for Logistic Regression is represented by the green curve, and the ROC analysis for Random Forest is represented by the orange curve. The graph shows that the Logistic Regression classifier performed better than the Random Forest classifier.

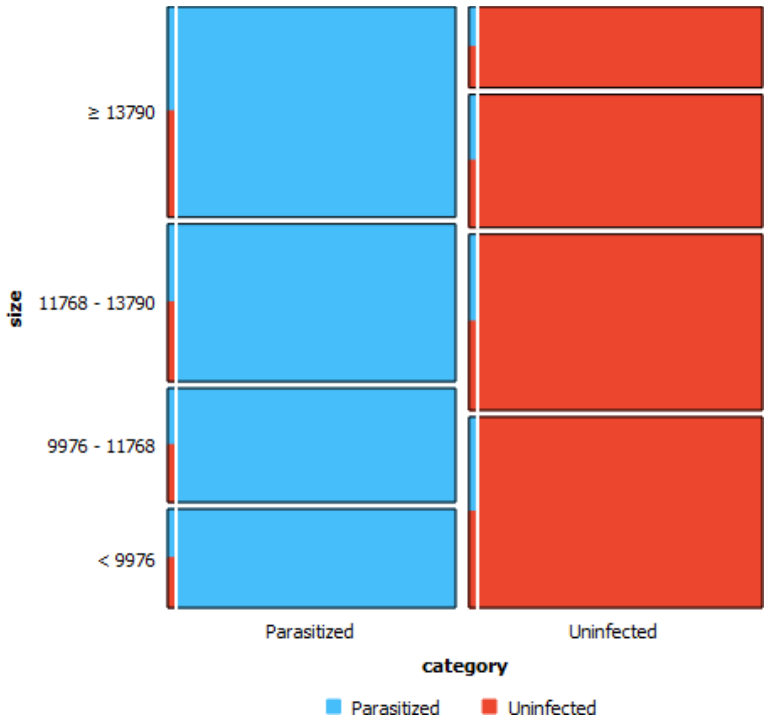
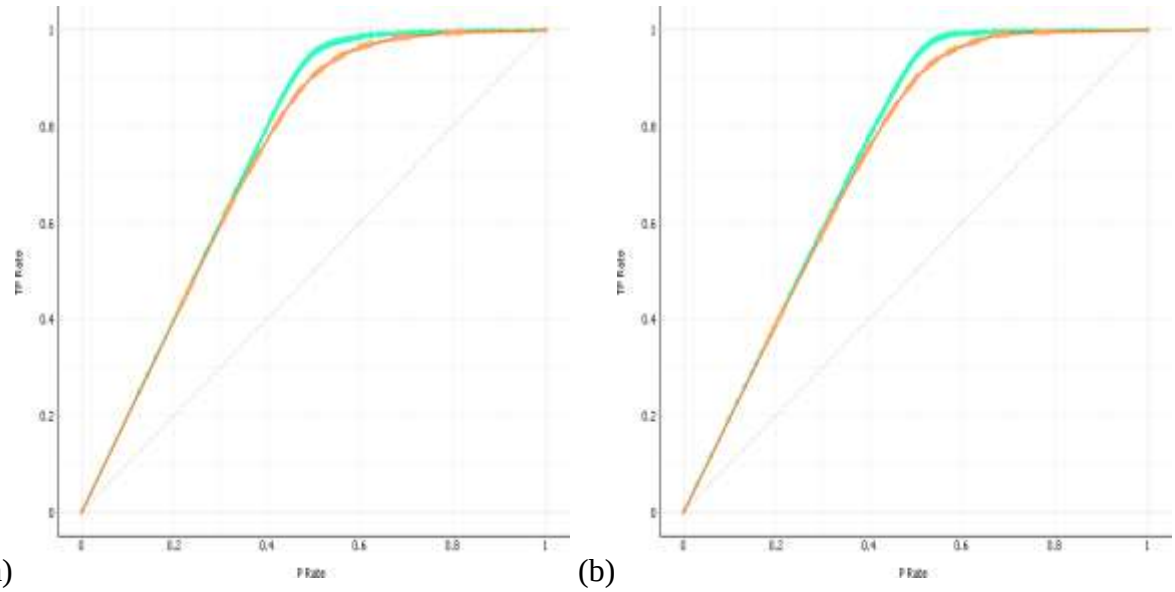


Figure 4.5: Mosaic display for two-way frequency table of the entire dataset.

Figure 4.5 shows a mosaic display for the malaria dataset, observing two variables (category and size) with an interior colouring as either parasitised or uninfected. The diagram shows that the category parasitised was mostly correctly predicted with few instances of incorrect prediction that is predicting parasitised images as uninfected images and also the category uninfected was mostly predicted with few instances of incorrect prediction that is predicting uninfected images as parasitised. The blue frequency represents parasitised images while red frequency represents uninfected images.



(a) (b)

Figure 4.6: Lift curve for (a) parasitized (b) uninfected images (whole dataset).

The relationship between the number of cases that were expected to be positive and those that really are positive is depicted by the graph in figure 4.6. The cumulative number of cases is plotted on the x-axis, while the cumulative number of true positives is plotted on the y-axis. We examined two different classifiers—Logistic Regression and Random Forest classifiers—as well as their performance against a random model were examined by sending them to a lift curve. The green line shows the performance of Logistic Regression, while the orange line shows that of Random Forest. It can be seen from the graph that Logistic Regression is the best classifier.

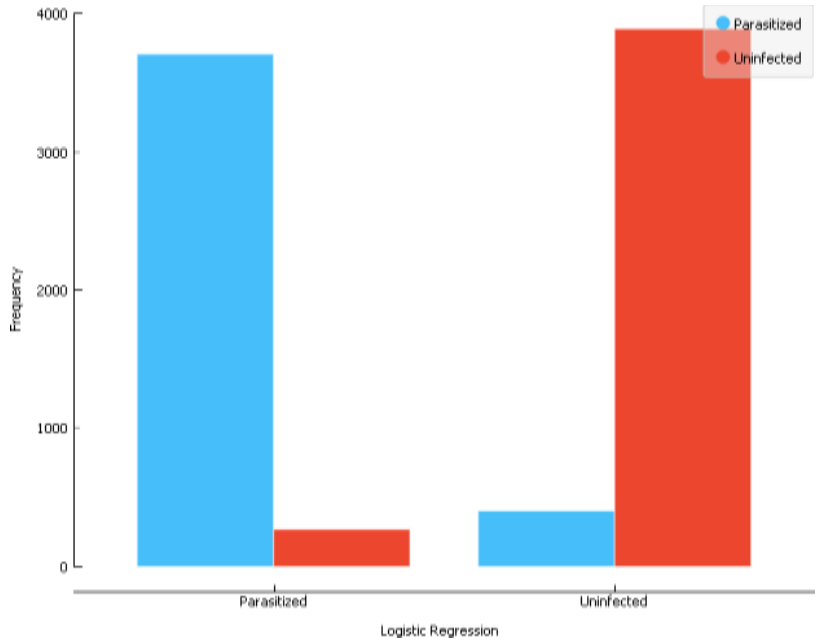


Figure 4.7: Frequency distribution for each attribute value in the whole dataset.

The graph in Figure 4.7 shows the frequency with which each attribute value appears in the dataset while using Logistic Regression classifier. The first bar shows the distributions for parasitised images, the bar in blue which shows higher frequency represents the parasitized images predictions that are actually correct, while the bar in red which shows very low frequency represents uninfected images that were predicted as parasitised. The second bar shows the distributions for uninfected images. The bar in blue which shows a lower frequency represents parasitised images that were predicted as uninfected, while the bar in red which shows a higher frequency represent the uninfected images predictions that are correct.

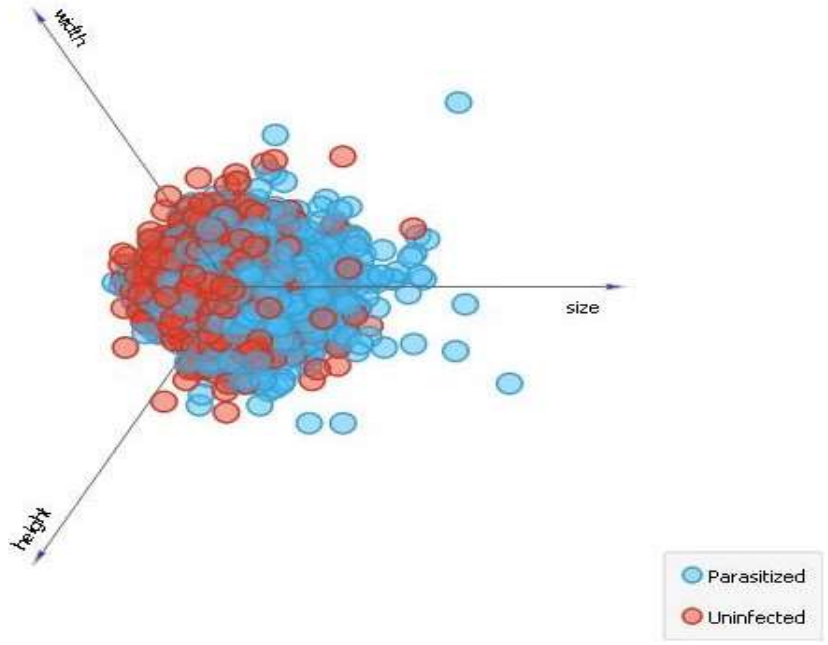


Figure 4.8: Linear projection showing a projection of the Malaria dataset.

From Figure 4.8 it was observed that width, size, and height are the best attributes separating parasitised images from uninfected images. The blue point represents parasitised images while the red points represent uninfected images.

Now reducing the images to 5000 and running the images to 5000 and running test and score, the accuracy of KNN is 90%, SVM is 79%, Random Forest is 99% and Logistic Regression returned an accuracy of 96%. These results are summarised in Table 4.2. By analysing this, we can see that Logistic Regression and Random Forest algorithms for this dataset, develop the most precise machine learning models.

Table 4.2: Evaluation Results using test and score for 5000 images

Model	AUC	CA	F1	Precision	Recall
Knn	0.976	0.908	0.908	0.912	0.908
SVM	0.926	0.796	0.794	0.816	0.796
Random forest	1.000	0.994	0.994	0.994	0.994
Logistic Regression	0.995	0.968	0.968	0.968	0.968

Due to their impressive performances, the models we are using for this study is the Logistic Regression and Random Forest. Using the remaining data to test our models, we have an illustration given in Figure 4.9.

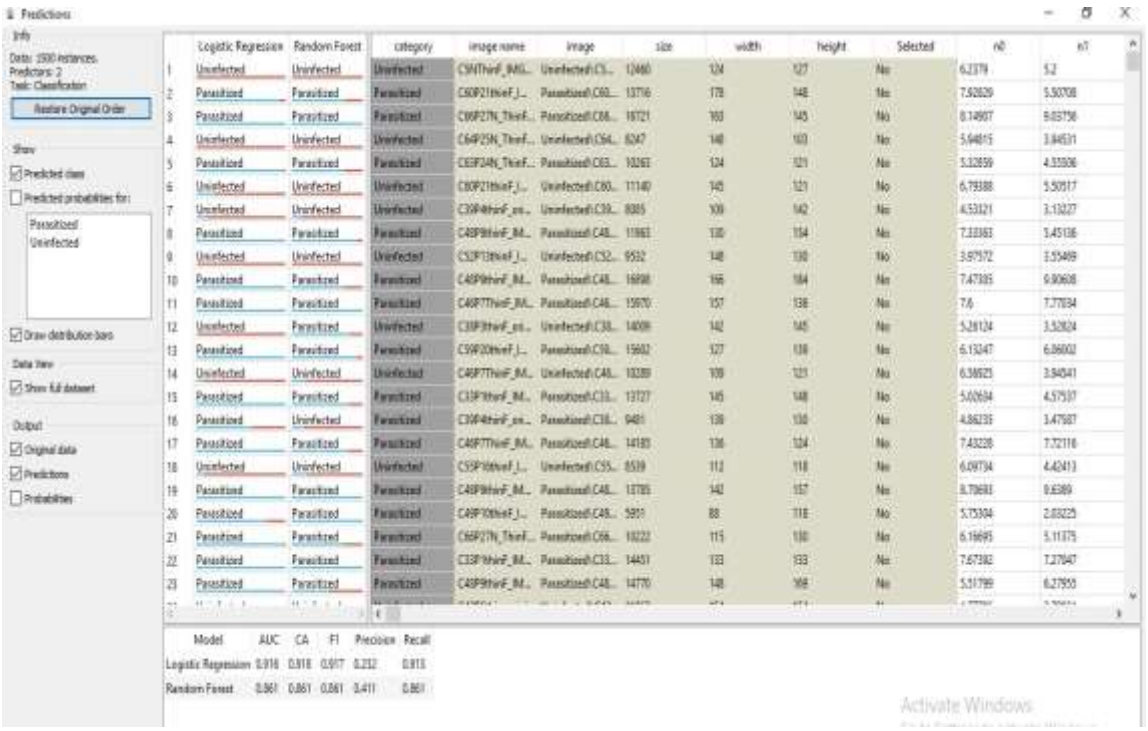


Figure 4.9: Prediction result of the test data (selected 5000 images) for logistic regression and random forest.

Figure 4.9 shows the prediction results of the test data using 30% of the remaining data, the data consist of 1500 instances, 2 predictors (Logistic Regression and Random Forest), from the prediction result it is observed that Logistic Regression and Random Forest classifier were able to predict correctly most parasitised and uninfected images, however there are few cases in which misclassification occurred.

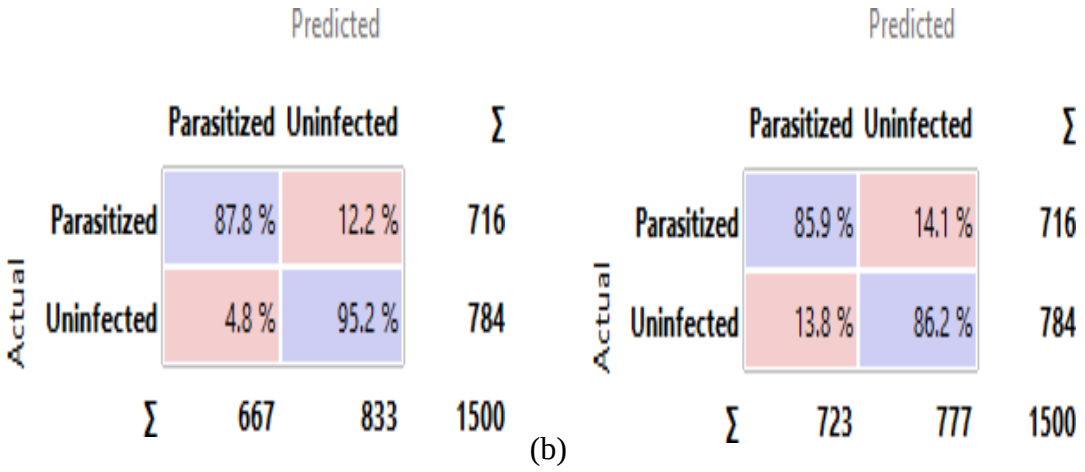


Figure 4.10: Predicted and actual parasitised red blood cells as shown in Confusion matrix of 5000 images using (a) logistic regression (b) Random Forest as classifier.

Confusion matrix reports on actual image classes and predicted classes and provides a data instance count for each combination. What was gotten right is highlighted with blue while the misclassified are reported with pink in Figure 4.10.

From the confusion matrix in Figure 4.10a, while using logistic regression as the classifier, it shows that 4.8% of the data was uninfected but was predicted as parasitised while 12.2% of the data that was parasitised was labelled as uninfected. However, 87.8% of parasitised images were correctly predicted and 95.2% of the uninfected images were correctly predicted while from Figure 4.10b, using random forest as the classifier, it shows that 13.8% of the data

was uninfected but was predicted as parasitised while 14.1% of the data that was parasitised was labelled as uninfected. However, 85.9% of parasitised images were correctly predicted and 86.2% of the uninfected images were correctly predicted.

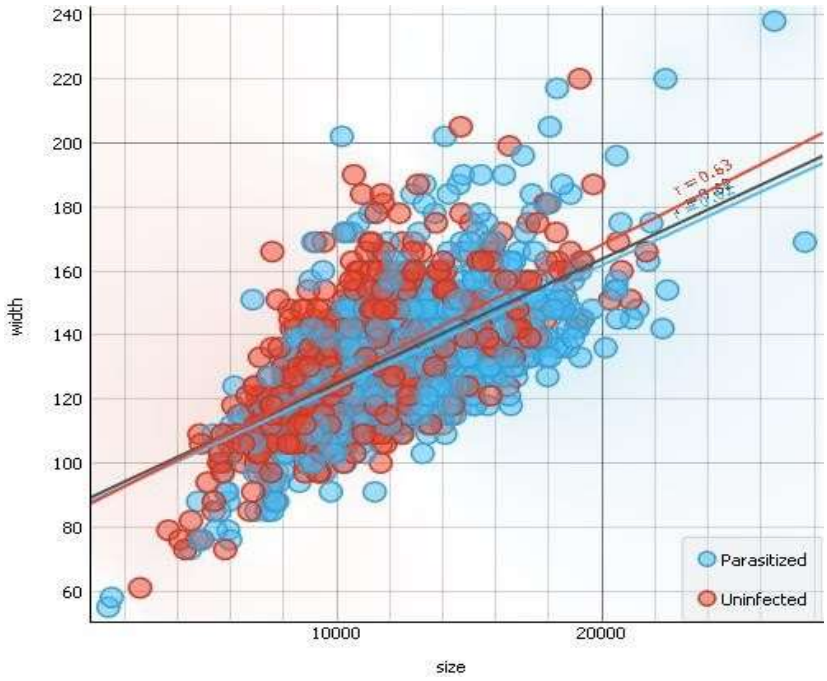
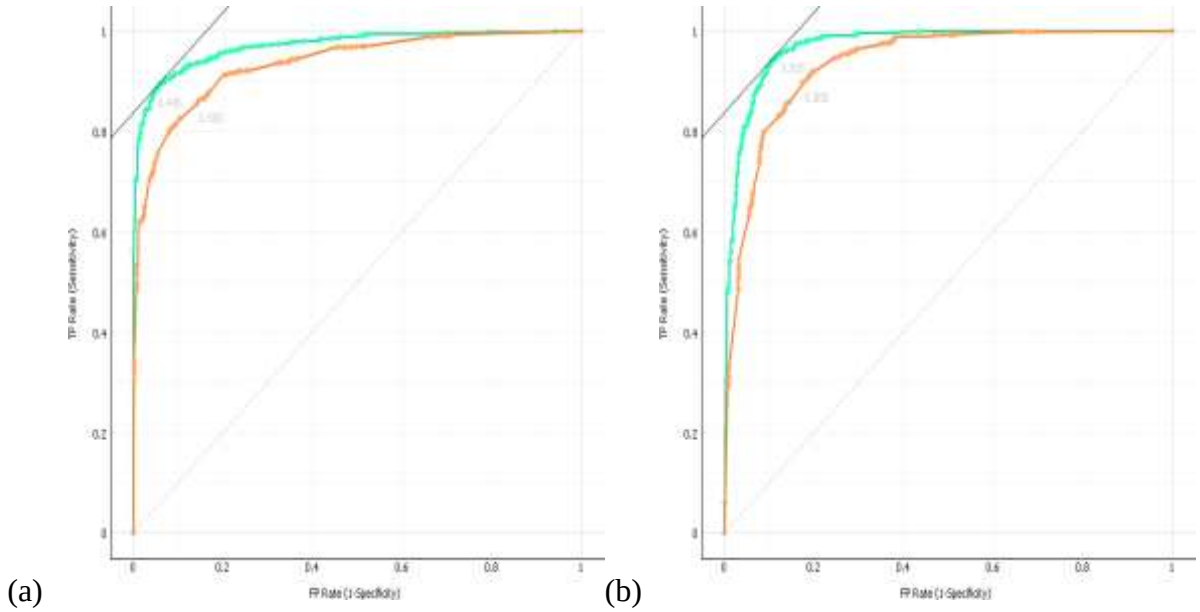


Figure 4.11: Scatter plot showing a 2 dimensional scatter plot visualisation for the 5000-image data.

Figure 4.11 displays a 2 dimensional scatter plot visualisation of the 5000 thin smear Geimsa-stained images, The data is represented as a set of points, where each point's size value on the x-axis determines its location on the horizontal axis and its width value on the y-axis determines its position on the vertical axis. The data points for parasitised images are represented in blue and data points for uninfected images are represented in red.



(a) (b)

Figure 4.12: ROC Analysis for (a) parasitised images (b) uninfected images showing plots of true positive rate against false positive rate for 5000-image data.

On the x-axis of Figure 4.12 (a,b) curve are graphs of the false positive rate (1-specificity; likelihood that the target is 1 when the true value is 0). sensitivity (probability that the goal equals 1 when the true value equals) is plotted against the true positive rate on the y-axis. The proximity of the curve to the top and left borders shows how precise the classifiers are. The ROC analysis for Logistic Regression is represented by the green curve, and the ROC analysis for Random Forest is represented by the orange curve. The graph shows that the Logistic Regression classifier performed better than the Random Forest classifier.

.

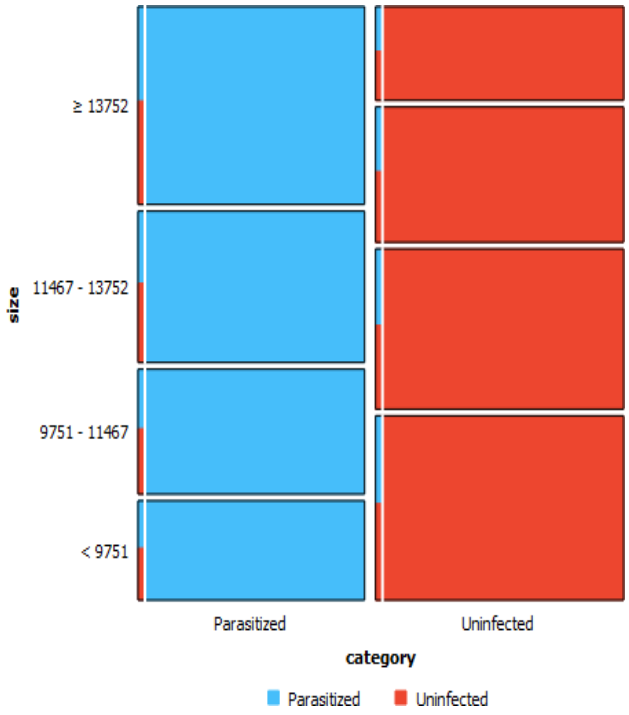
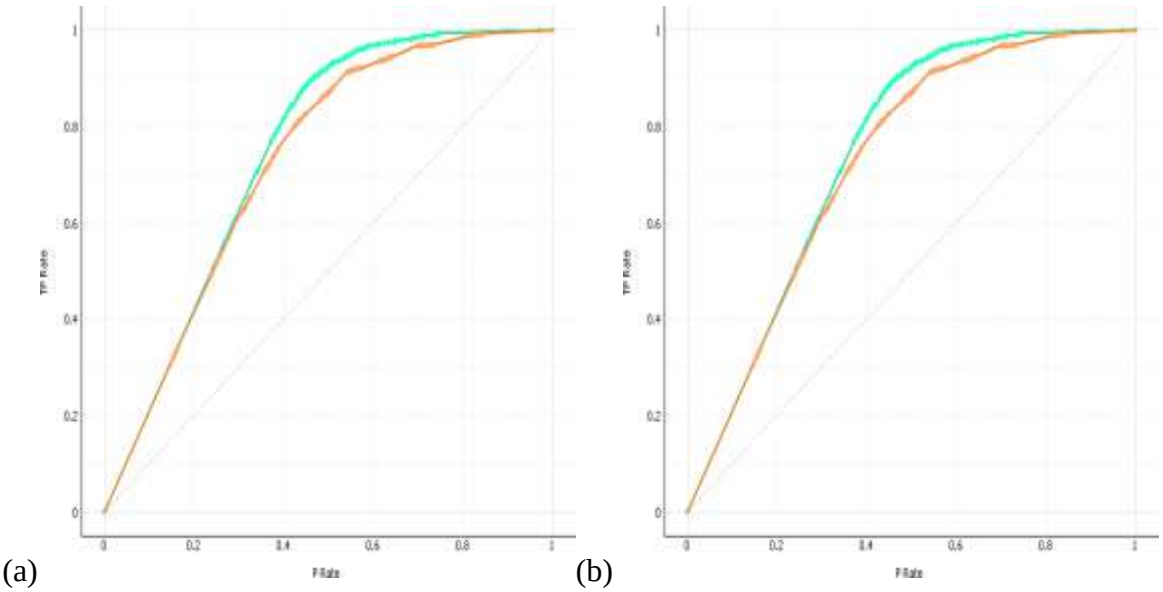


Figure 4.13: Mosaic display showing a two-way frequency table for the 5000-image data.

Figure 4.13 shows a mosaic display for the malaria dataset, observing two variables (category and size) with an interior colouring as either parasitised or uninfected. The diagram shows that the category parasitised was mostly correctly predicted with few instances of incorrect prediction that was predicting parasitised images as uninfected images and also the category uninfected was mostly predicted with few instances of incorrect prediction that was predicting uninfected images as parasitised. The blue frequency represents parasitised images while red frequency represents uninfected images.



(a) (b)

Figure 4.14: Lift curve for (a) parasitised (b) uninfected images (5000-image data) showing the relation between predicted positive and actual positive.

The relationship between the number of cases that were expected to be positive and those that really are positive is depicted by the graph in Figure 4.14. The cumulative number of cases is plotted on the x-axis, while the cumulative number of true positives is plotted on the y-axis. Two different classifiers—Logistic Regression and Random Forest classifiers—as well as their performance against a random model were examined by sending them to a lift curve. The green line shows the performance of Logistic Regression, while the orange line shows that of Random Forest. It can be seen from the graph that Logistic Regression is the best classifier.

.

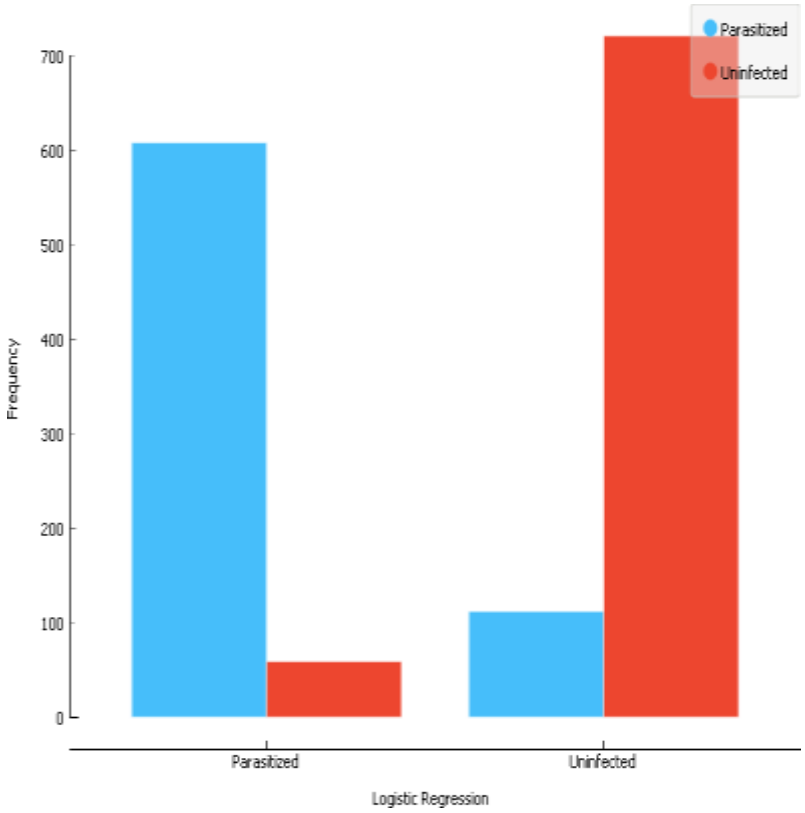


Figure 4.15: Frequency distribution for each attribute value appears in the 5000-image data.

The graph in Figure 4.15 shows how many times each attribute value appears in the dataset while using Logistic Regression classifier. The first bar showed the distributions for parasitised images, the bar in blue which shows higher frequency represents the parasitized images predictions that are correct, while the bar in red which shows very low frequency represents uninfected images that were predicted as parasitised. The second bar showed the distributions for uninfected images. The bar in blue which shows a lower frequency represents parasitized images that were predicted as uninfected, while the bar in red which shows a higher frequency represent the uninfected images predictions that are correct.

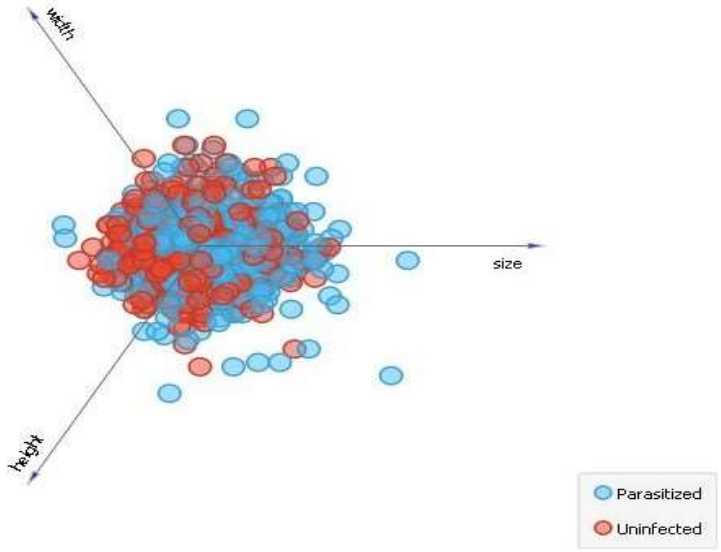


Figure 4.16: Linear projection showing a projection of the Malaria dataset (5000-image data).

From Figure 4.16 it is observed that width, size and height are the best attributes separating parasitised images from uninfected images. The blue points represent parasitised images while the red points represent uninfected images.

4.2 Unsupervised Learning

4.2.1 Hierarchical clustering

To put the data into logical groups, hierarchical clustering was used to discover groups or subgroups, the images distance was used to create hierarchical clustering, Dendograms, which are trees that show the structure of the identified clusters and the separation between them, are displayed because of hierarchical clustering. To make the dendogram more telling. The image viewer was connected to the dendogram to see images in each cluster.

It identifies clusters that are closest to each other and merge the most similar clusters as illustrated in figure 4.17.

Figure 4.17: Hierarchical clustering displaying the dendogram. Hierarchical clustering was established using the cosine distance.

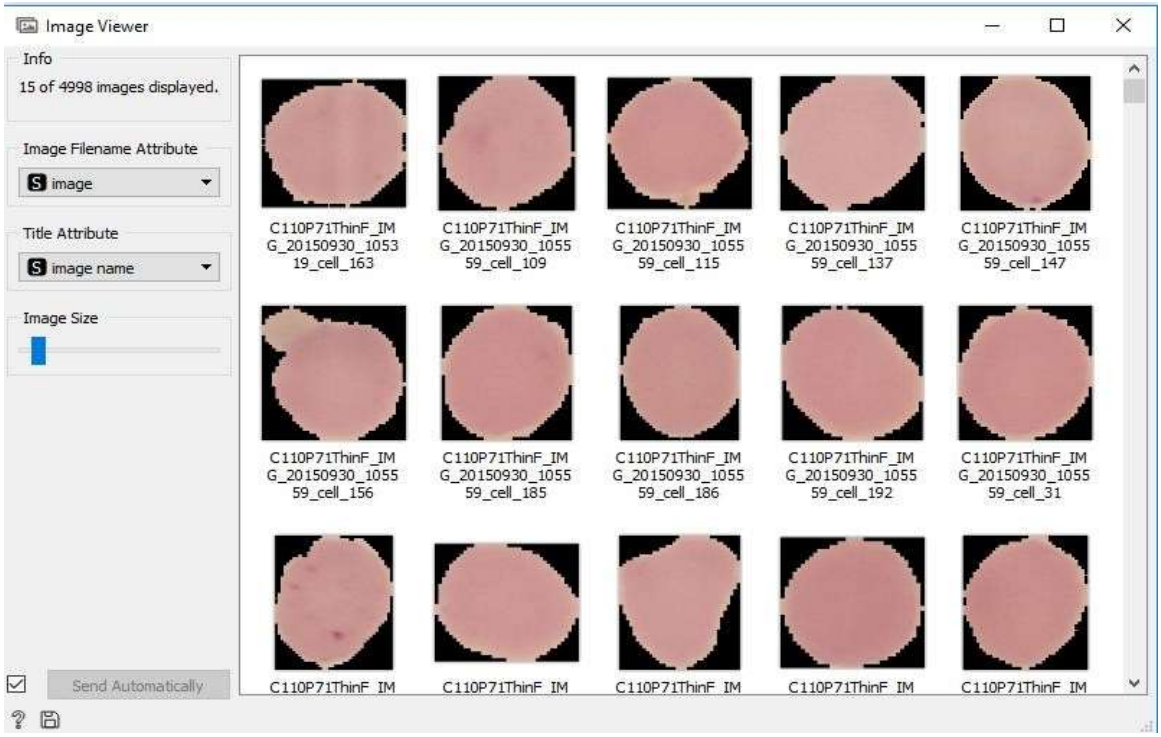


Figure 4.18: Image viewer showing uninfected images clustered together.

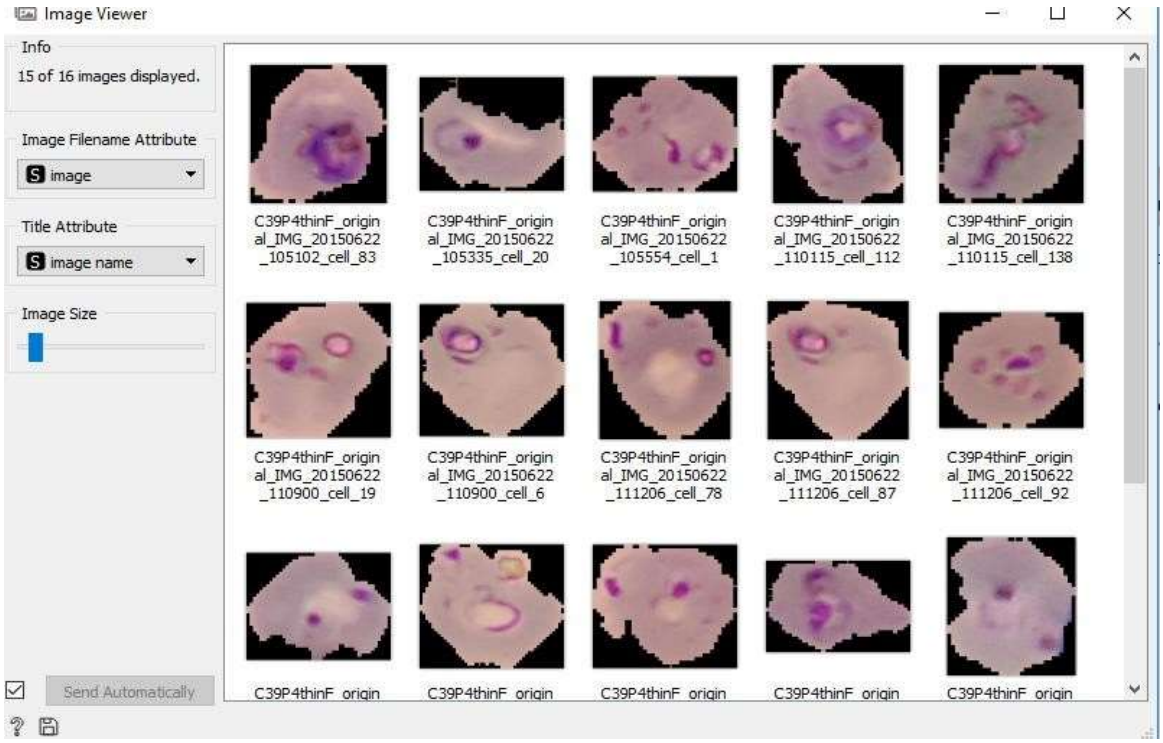


Figure 4.19: Image viewer showing infected images clustered together.

Figure 4.17 shows a hierarchical cluster analysis displaying a dendogram**,** Hierarchical clustering was established using the cosine distance. It grouped similar images from the malaria datasets into groups known as clusters. The dendogram shows several clusters, each of which is distinct from the others and contains objects that are generally like one another. Figure 4.18 shows the results from selecting a cluster, it is observed that all the images in that cluster are like each other as they are all parasitised images. Figure 4.19 shows the results from selecting another cluster, it is also observed that all images are homogenous and heterogeneous from the images in the other cluster. This shows that the clusters indeed make sense as images that are parasitised are clustered together while those that are uninfected are also clustered together.

4.2.2 k-means algorithm implementation

k-means algorithm is often used to find interesting groups of data instances such as segmentation of customers based on their shopping habit, finding similar documents, or grouping tweets based on the contents. k-means can also be used to find clusters, the k-means discovered two clusters as expected. To confirm that silhouette score was used, and it gave the choice of two clusters as the best.

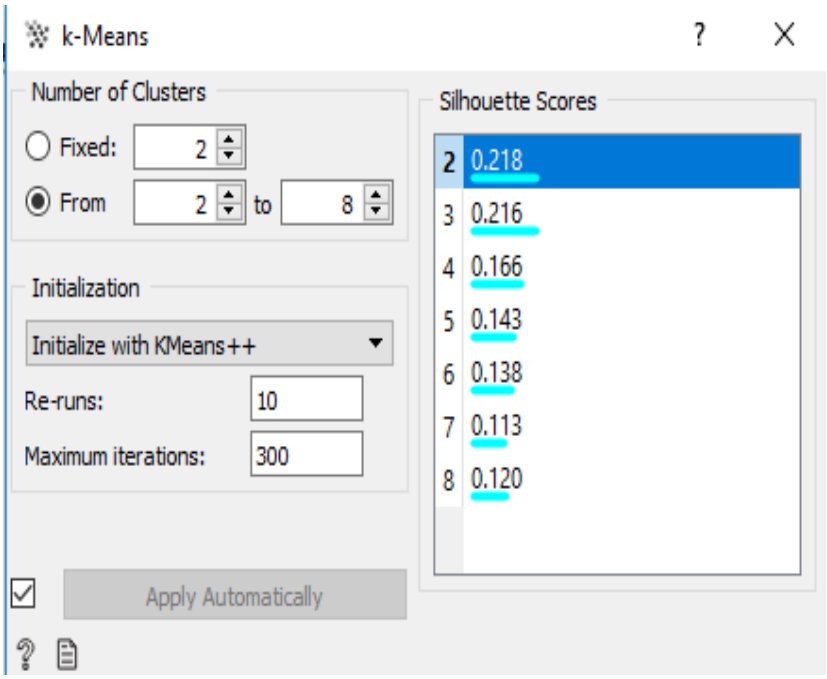


Figure 4.20: Silhouette scores giving choice of two clusters.

From Figure 4.20, k-means algorithm was implemented, to evaluate the result, the best choice of number of clusters for the dataset was investigated using k-means, a choice of the best number of clusters from 2 to 8 was requested and the Silhouette scores gave a choice of two clusters, which is the expected result since we are grouping the images into two groups that is either parasitised or uninfected. To visualize this implementation, k means could be connected to any visualization. A scatter plot for this implementation is given in Figure 4.21.

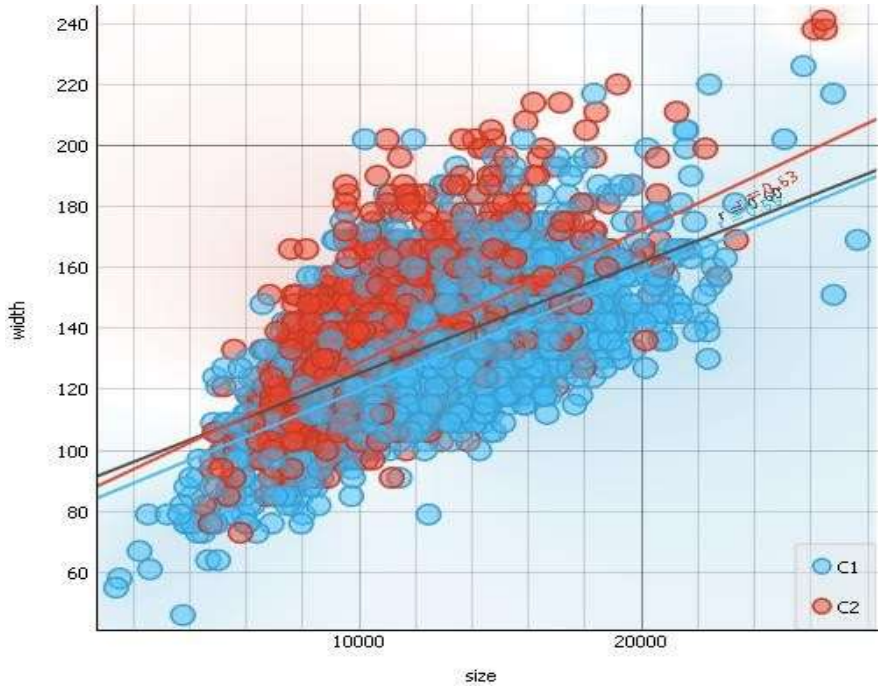


Figure 4.21: Scatter plot showing the two clusters.

In Figure 4.21, a two-dimensional scatter plot visualization of the k-means algorithm is shown. The data is represented as a set of points, with each point's size value on the x-axis determining its position on the horizontal axis and its width value on the y-axis determining its position on the vertical axis. Due to the usage of an unlabelled dataset and the algorithm's inability to determine the classes of the photos, the data points were represented as clusters 1 and 2, respectively.

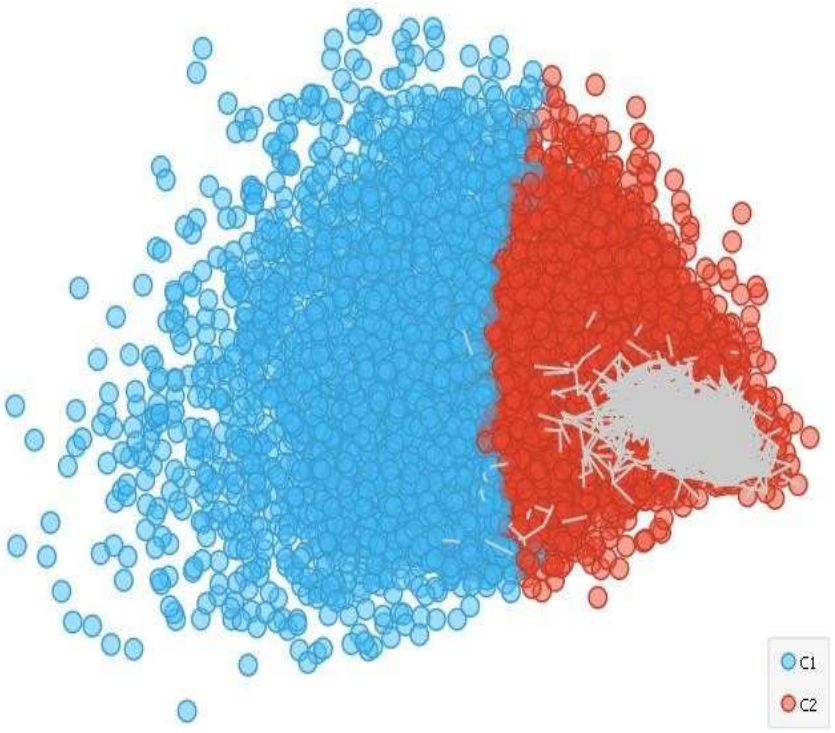


Figure 4.22: Multidimensional scaling (MDS) showing two-dimensional projection of points.

Figure 4.22 shows a multidimensional scaling (MDS) displaying a two-dimensional projection of points, it iteratively moves around in a simulation of a physical model, there is a force which push two points that are too close to each other together and pulling points that are too far apart away. The data points are represented as cluster 1 and cluster 2 since unlabelled dataset was used and the algorithm have no way of knowing the classes of the images.

CHAPTER FIVE

5.0 CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The aim of this work was to contribute to the development of supervised and unsupervised machine learning algorithms for detecting malaria parasites in thin blood smear images using Orange software. Through review of previous work on this subject, the use of neural networks was made as a promising technique for automated image interpretation.

Two methods of machine learning algorithm were used, supervised and unsupervised. For supervised learning, four classifiers were used which include Logistic Regression, Random Forest, KNN and SVM. These classifiers were used to train 70% of the datasets, a cross validation of the dataset was done and an accuracy of 95% for logistic regression, 99% for random forest, 92% for KNN and 78% for SVM was obtained. Since logistic regression and random forest had the highest accuracy, they were used to test the remaining 30% of the data.

Using logistic regression as the classifier, it shows that 3.5% of the data was uninfected but was predicted as parasitised (False Positive) while 6.5% of the data that was parasitised was labelled as uninfected (False Negative). However, 93.5% of parasitised images were correctly predicted (True Positive) and 96.5% of the uninfected images were correctly predicted (True Negative).

Using random forest as the classifier, it showed that 9.5% of the data was actually uninfected but was predicted as parasitised (False Positive) while 9.6% of the data that was actually

parasitised was labelled as uninfected (False Negative). However, 90.5% of parasitised images were correctly predicted (True Positive) and 90.4% of the uninfected images were correctly predicted (True Negative). A sensitivity and specificity of 96% and 94% respectively was achieved with Logistic regression classifier while a sensitivity and specificity of 90.4% and 90.5% was achieved with Random Forest Classifier. All the classifiers correctly predicted more than 90%.

For unsupervised learning, Hierarchical clustering and k-means was used to cluster similar images together, hierarchical clustering was able to group parasitised images in one cluster and uninfected in another cluster although there were few instances where parasitised and uninfected were clustered together. While k-means also discovered two clusters from the data sets.

As a result, it is concluded that the developed machine learning algorithm cannot entirely replace the requirement for experienced professionals in the interpretation of thin blood smears for malaria diagnosis. But greater than 90% accuracy in automatic determination is a major step in the right direction. Furthermore, when the results of the classification are presented to an expert in the visual way that was shown here, this expert can easily determine the true infection status of the objects predicted as infected. This would greatly reduce the number of cells that need to be evaluated. It is therefore believed that the method can contribute to reducing the diagnostic burden and increasing the availability of malaria diagnostics globally.

5.2 Recommendations

The following recommendations are made:

i. This study's focus was only on the interpretation of P. Falciparum Giemsa-stained thin blood smears. This was chosen, as it is commonly the most widely accepted technique for the microscopic diagnosis of malaria, as well as the diagnostic method most limited by the time consumed to interpret the data. The choice to use only P. Falciparum infected samples was made based on availability, and because this is the most predominant species and most deadly species in the world. Since no attempts have been made to automatically distinguish between different species in this experiment, we strongly suggest that this be a focus of future study.

ii. Future study in this field is highly encouraged. To automatically interpret blood films, a modest computing device or mobile application platform that runs the algorithms can be integrated into the microscope design.

iii. Additionally, fluorescence microscopy is another malaria microscopy technique that is already used and could benefit from automation. However, this was not tested because there was no such data available; as a result, it is suggested that this be the subject of future research.

5.3 Contribution to Knowledge

This study established that supervised and unsupervised machine learning algorithms for detection of malaria parasites in thin blood smears. Using 27558 thin Giemsa-stained images from the National Institute of Health, USA, k-nearest neighbour (kNN), support vector machine (SVM), random forest (RF) and logistic regression (LR) were employed in training the data. Classification accuracy of 90.8%, 78.8%, 99.3% and 95.5% for kNN, SVM, RF and

LR respectively. Hierarchical and k-means clustering algorithms were used for unsupervised model training of the dataset. Silhouette score of 0.218 was found for two clusters during implementation of k-means clustering. These results showed that random forest algorithm produced the best classification results when malaria disease state is known while k-means clustering performed well for cases in which malaria disease state is unknown. This algorithm when integrated into the microscope design can automatically detect malaria parasite, thus, give automatic interpretation of images for malaria diseases. That has several advantages compared with manual diagnosis, such as providing a more reliable interpretation of blood films, allowing more patients to be attended with more precision and accuracy, finally it would leads to the reduction in diagnostic costs.

REFERENCES

Acharya, U. R., Molinari, F., Sree, S. V., Chattopadhyay, S., Ng, K. H., & Suri, J. S. (2012). Automated diagnosis of epileptic EEG using entropies. *Biomedical Signal Processing and Control*, *7*(4), 401-408.

Adeoye, G. O., & Nga, I. C. (2007). Comparison of Quantitative Buffy Coat technique (QBC) with Giemsa-stained Thick Film (GTF) for diagnosis of malaria. *Parasitology international*, *56*(4), 308-312.

Anggraini, D., Nugroho, A. S., Pratama, C., Rozi, I. E., Pragesjvara, V., & Gunawan, M. (2011). Automated status identification of microscopic images obtained from malaria thin blood smears using Bayes decision: a study case in Plasmodium falciparum. In *2011 International Conference on Advanced Computer Science and Information Systems* (pp. 347-352). IEEE.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Center for Disease Control and Prevention. (2013). *Comparison of the Plasmodium Species Which Cause Human Malaria.* Center for Disease Control and Prevention. Atlanta, USA.

Center for Disease Control and Prevention. (2018). *CDC Parasites—Malaria*. Center for Disease Control and Prevention. Atlanta, USA.

Chen, T., & Chen, H. (1995). Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, *6*(4), 911-917.

Das, D. K., Maiti, A. K., & Chakraborty, C. (2015). Automated system for characterization and classification of malaria-infected stages using light microscopic images of thin blood smears. *Journal of microscopy*, *257*(3), 238-252.

Delgado-Ortet, M., Molina, A., Alférez, S., Rodellar, J., & Merino, A. (2020). A deep learning approach for segmentation of red blood cell images and malaria detection. *Entropy*, *22*(6), 657.

Devi, S. S., Sheikh, S. A., & Laskar, R. H. (2016). Erythrocyte Features for Malaria Parasite Detection in Microscopic Images of Thin Blood Smear: A Review. *Int. J. Interact. Multim. Artif. Intell.*, *4*(2), 34-39.

Dong, Y., Jiang, Z., Shen, H., Pan, W. D., Williams, L. A., Reddy, V. V., Benjamin, W. H., & Bryan, A. W. (2017, February). Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells. In *2017 IEEE EMBS international conference on biomedical & health informatics (BHI)* (pp. 101-104). IEEE.

Edgar, T., & Manz, D. (2017). *Research methods for cyber security*. Syngress. Oxford, UK.

El Bouchefry, K., & de Souza, R. S. (2020). Learning in big data: Introduction to machine learning. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (pp. 225-249). Elsevier.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256). JMLR Workshop and Conference Proceedings.

Gopakumar, G. P., Swetha, M., Sai Siva, G., & Sai Subrahmanyam, G. R. K. (2018). Convolutional neural network-based malaria diagnosis from focus stack of blood smear images acquired using custom-built slide scanner. *Journal of biophotonics*, *11*(3), e201700003.

Grabias, B., & Kumar, S. (2016). Adverse neuropsychiatric effects of antimalarial drugs. *Expert opinion on Drug safety*, *15*(7), 903-910.

Haykin, S. (1998). Neural Networks: A Comprehensive Foundation Upper. *Saddle River. NJ, USA*, 1-842.

Houwen, B. (2002). Blood film preparation and staining procedures. *Clinics in laboratory medicine*, *22*(1), 1-14.

Jan, Z., Khan, A., Sajjad, M., Muhammad, K., Rho, S., & Mehmood, I. (2018). A review on automated diagnosis of malaria parasite in microscopic blood smears images. *Multimedia Tools and Applications*, *77*(8), 9801-9826.

Janse, C. J., & Van Vianen, P. H. (1994). Flow cytometry in malaria detection. *Methods in cell biology*, *42*, 295-318.

Kawamoto, F. (1991). Rapid diagnosis of malaria by fluorescence microscopy with light microscope and interference filter. *The Lancet*, *337*(8735), 200-202.

Keiser, J., Utzinger, J., Premji, Z., Yamagata, Y., & Singer, B. H. (2002). Acridine Orange for malaria diagnosis: its diagnostic performance, its promotion and implementation in Tanzania, and the implications for malaria control. *Annals of Tropical Medicine & Parasitology*, *96*(7), 643-654.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Lam, E. Y. (2005, June). Combining gray world and retinex theory for automatic white balance in digital photography. In *Proceedings of the Ninth International Symposium on Consumer Electronics, 2005. (ISCE 2005).* (pp. 134-139). IEEE.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Linder, N., Turkki, R., Walliander, M., Mårtensson, A., Diwan, V., Rahtu, E., Pietikäinen, M., Lundin, M., & Lundin, J. (2014). A malaria diagnostic tool based on computer vision screening and visualization of Plasmodium falciparum candidate areas in digitized blood smears. *PLoS One*, *9*(8), e104855.

Ling, S. J., Sanny, J., & Moebs, W. (2016). Microscopes and Telescopes. *University Physics Volume, Openstax. Houston, Texas.*

Mehanian, C., Jaiswal, M., Delahunt, C., Thompson, C., Horning, M., Hu, L., McGuire, S., Ostbye, T., Mehanian, M., Wilson, B., Champlin, C., Long, E., Proux, S., Gamboa, D., Chiodini, P., Carter, J., Dhorda, M., Isaboke, D., Ogutu, B., Oyibo, W., Villasis, E., Tun, K. M., Bachman, C., & Bell, D. (2017). Computer-automated malaria diagnosis and quantitation using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 116-125).

Mitiku, K., Mengistu, G., & Gelaw, B. (2003). The reliability of blood film examination for malaria at the peripheral health unit. *Ethiopian Journal of Health Development*, *17*(3), 197-204.

Mushabe, M. C., Dendere, R., & Douglas, T. S. (2013). Automated detection of malaria in Giemsa-stained thin blood smears. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3698-3701). IEEE.

Mustafa, W. A., Alquran, H., Aihsan, M. Z., Saifizi, M., Khairunizam, W., Abdul-Nasir, A. S., & Nasrudin, M. W. (2021, November). Malaria Parasite Diagnosis Using Computational Techniques: A Comprehensive Review. In *Journal of Physics: Conference Series* (Vol. 2107, No. 1, p. 012031). IOP Publishing.

Nasir, A. A., Mashor, M. Y., & Mohamed, Z. (2012, December). Segmentation based approach for detection of malaria parasites using moving k-means clustering. In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences* (pp. 653-658). IEEE.

Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., & Thoma, G. (2018). Image analysis and machine learning for detecting malaria. *Translational Research*, *194*, 36-55.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., & Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, *6*, e4568.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Ross, N. E., Pritchard, C. J., Rubin, D. M., & Duse, A. G. (2006). Automated image processing method for the diagnosis and classification of malaria on thin blood smears. *Medical and Biological Engineering and Computing*, *44*(5), 427-436.

Savkare, S. S., Narote, A. S., & Narote, S. P. (2016). Automatic blood cell segmentation using K-Mean clustering from microscopic thin blood images. In *Proceedings of the Third International Symposium on Computer Vision and the Internet* (pp. 8-11).

Sharif, J. M., Miswan, M. F., Ngadi, M. A., Salam, M. S. H., & bin Abdul Jamil, M. M. (2012, February). Red blood cell segmentation using masking and watershed algorithm: A preliminary study. In *2012 International Conference on Biomedical Engineering (ICoBE)* (pp. 258-262). IEEE.

Shillcutt, S., Morel, C., Goodman, C., Coleman, P., Bell, D., Whitty, C. J., & Mills, A. (2008). Cost-effectiveness of malaria diagnostic methods in sub-Saharan Africa in an era of combination therapy. *Bulletin of the World Health Organization*, *86*, 101-110.

Shute, G. T., & Sodeman, T. M. (1973). Identification of malaria parasites by fluorescence microscopy and acridine orange staining. *Bulletin of the World Health Organization*, *48*(5), 591.

Silamut, K., & White, N. J. (1993). Relation of the stage of parasite development in the peripheral blood to prognosis in severe falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *87*(4), 436-443.

Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. *Cambridge University, UK*, *32*(34), 2008.

Suwalka, I., Sanadhya, A., Mathur, A., & Chouhan, M. S. (2012). Identify malaria parasite using pattern recognition technique. In *2012 International Conference on Computing, Communication and Applications* (pp. 1-4). IEEE.

Tangpukdee, N., Duangdee, C., Wilairatana, P., & Krudsood, S. (2009). Malaria diagnosis: a brief review. *The Korean journal of parasitology*, *47*(2), 93.

Tek, F. B., Dempster, A. G., & Kale, I. (2010). Parasite detection and identification for automated thin blood film malaria diagnosis. *Computer vision and image understanding*, *114*(1), 21-32.

van Driel, N. (2020). Automating malaria diagnosis: a machine learning approach: Erythrocyte segmentation and parasite identification in thin blood smear microscopy images using convolutional neural networks. *Delft center for systems and control*, pp. 10-15.

Vink, J. P., Laubscher, M., Vlutters, R., Silamut, K., Maude, R. J., Hasan, M. U., & De Haan, G. (2013). An automatic vision-based malaria diagnosis system. *Journal of microscopy*, *250*(3), 166-178.

World Health Organization. (2016). *Malaria microscopy quality assurance manual-version 2*. World Health Organization. Geneva, Switzerland.

World Health Organization. (2017). *Malaria microscopy quality assurance manual-version 2*. World Health Organization. Geneva, Switzerland.

World Health Organization. (2018). *High burden to high impact: a targeted malaria response* (No. WHO/CDS/GMP/2018.25 Rev. 1). World Health Organization.

World Health Organization. (2019). *Compendium of WHO malaria guidance: prevention, diagnosis, treatment, surveillance and elimination* (No. WHO/CDS/GMP/2019.03). World Health Organization.

World Health Organization. (2021). WHO malaria policy advisory group (MPAG) meeting: meeting report, April 2021.

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition* (pp. 2528-2535). IEEE.

Zou, L. H., Chen, J., Zhang, J., & Garcia, N. (2010). Malaria cell counting diagnosis within large field of view. In *2010 International Conference on Digital Image Computing: Techniques and Applications* (pp. 172-177). IEEE.