
# FYP17028

# Classification for Pathological Images Using Machine Learning

Intermediate Report

Chi Ian Tang

3035209241

Supervisor: Dr. S. M. Yiu

Department of Computer Science

The University of Hong Kong

21 January 2018

# Abstract

Diagnostic microscopy is currently used for the diagnosis of many common infections including bacterial vaginosis and malaria, but the dependence of diagnostic microscopy on human expertise limits its availability. Recent attempts of using machine learning algorithms in the development of automated diagnostic tools have been successful. In this project, I developed a diagnostic system for bacterial vaginosis using a number of computer vision and machine learning algorithms. The first prototype developed has an accuracy of 45.1% in correctly identifying the degree of infection, and an accuracy 9.6% in correctly identifying the exact Nugent Score. The next prototype is currently under development, with changes tailored to tackle the problems in the first prototype and aimed to improve performance. The project will then go on to the development of the final diagnostic system, which is expected to have desirable performance and can be used as the starting point for developing a robust and professional diagnostic system. The development processes will be analysed after the development process, such that this project can act as an example for future attempts in solving similar problems.

# Acknowledgment

I would like to express my sincere gratitude towards Dr. S. M. Yiu, the supervisor of the project, for his time and effort in guiding me in the entire project, as well as giving me timely feedback on my work. I would also like to express my gratitude towards Mr. Keith Chau, the course instructor for CAES9542 Technical English for Computer Science subclass F, for his teaching and help in the preparation of the project plan as well as the intermediate report. I would like to express my appreciation to Dr. Dirk Schnieders, for his help and professional suggestions in the aspect of computer graphics, and Ms. Bingbin Liu, who spent much time in the setting up the framework and building the first prototype of the diagnostic tool. In addition, I would like to express my gratitude towards Professor Patrick C. Y. Woo, for his professional feedback given in the perspective of a medical professional, and his help in the arrangements in the acquisition of blood smear images. I also want to express my appreciation to Mr. Chris Chi-Ching Tsang, for his help in arranging meetings with the medical experts as well as his help in facilitating communication between different parties, and also to all the medical experts for their help in the laborious data labelling tasks.

# Table of Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| BV | Bacterial vaginosis |
| DBSCAN | Density-based spatial clustering of applications with noise |
| HIV | Human immunodeficiency virus |
| MSER | Maximally stable extremal regions |
| SQL | Structured Query Language |

# 1. Introduction

## 1.1. Background

Bacterial infection is a common medical condition in humans, and several pathogens were found to be responsible for the development of malignant tumours [1]. Bacterial vaginosis (BV), one of the most common bacterial infections in the vagina, was estimated to affect tens of millions of people in the United States of America alone [2]. The prevalence of this infection varies by countries and can be as high as 50% in women at reproductive age [1]. Studies have also shown that this infection increases the risks of being infected with human immunodeficiency virus (HIV) [1], [3]. The Nugent Score System [4], which involves the investigation of Gram-stained vaginal smears from patients, is considered to be the Gold Standard in diagnosing bacterial vaginosis [1], [5], [6]. In addition, diagnostic microscopy is also the main diagnostic method for parasitic infections, including Malaria, in major hospitals [7], [8].

## 1.2. Motivation

Diagnostic microscopy, however, requires a considerable amount of training and skills, where the accuracy often depends on how experienced the microscopist is [6], [9]. Furthermore, it could be time-consuming since it involves human diagnosis, and hence could be expensive for patients. In the light of the prevalence and consequences of aforementioned infections, an automated process could reduce the dependency on human expertise and provide a more affordable way to perform diagnosis.

Attempts in applying machine learning techniques to the diagnosis of several common infections have been successful with a high level of performance [10]. However, an automated system for the diagnosis of bacterial vaginosis based on patients' blood smear is still not present, and in many cases, the effects on a different number of training samples, the performances of a variety of machine learning models on the same set of data, etc. are not thoroughly discussed.

## 1.3. Objectives

This project aims to explore the possibilities in employing machine learning and computer vision techniques in diagnostic microscopy. An integrated, automated diagnosis system for bacterial vaginosis will be developed, such that the time and cost of bacterial vaginosis diagnosis could be reduced.

In addition, general strategies for, as well as limitations of applying machine learning and computer vision techniques in medical contexts will be explored, such that this project could be used as a guideline for future projects using similar techniques.

### 1.4. Scope

The project consists of two main components. First, an automated diagnostic tool which estimates the degree of infection based on a blood smear image, with a simple interface will be developed. The auxiliary data collection tool, which facilitates the collection of detailed information of blood smear images, will also be developed. This project does not directly involve the acquisition of blood smear images from patients nor the labelling process, where data are obtained from the medical experts from the Li Ka Shing Faculty of Medicine, the University of Hong Kong, and other publicly available sources. Second, a report on limitations and general strategies for applying machine learning techniques in diagnostic microscopy will be produced, which includes the limits on the number of training samples required and time required to train the machine learning models.

### 1.5. Related Works

A number of recent studies made use of a range of computer vision and machine learning techniques on diagnostic microscopy. In particular, Quinn et al. [10] explored the use of convolutional neural networks (a machine learning algorithm) in detecting several infections including tuberculosis and hookworm and the detection tools were very successful with high accuracy. Kraus et al. [11] combined convolutional neural networks and image segmentation with multiple instance learning in classifying segmented images only using high-level annotations for the entire image. These studies showed that deep learning techniques had a range of advantages in diagnostic microscopy and saw significant improvements over traditional techniques.

### 1.6. Outline of the report

The remainder of this report starts by outlining the major deliverables of the project, followed by the methodologies employed in the development of the deliverables. The current progress including the preliminary results and the current direction of the project is then given. Major difficulties encountered and mitigation strategies in the future development are elaborated in the next section, followed by a conclusion at the end.

# 2. Outline of Deliverables

## 2.1. Data collection tool

A simple auxiliary tool, *Clicklable* (see Section 3.3.3 for detailed information), for labelling the microscopic images was developed to facilitate detailed data labelling. Data labelling is an essential step for supervised machine learning, where known truths about input data are annotated with the expected output and these input / output pairs are then used to train machine learning models. This tool facilitates this process by allowing the user to load a blood smear image and perform labelling by clicking the locations where bacteria are present. The labelled points will be marked with a shape around it, and high level of customization can be done. The aim is to reduce the time required for labelling the images, as well as to tailor the data representation.

## 2.2. Automated diagnostic system for bacterial vaginosis

The main objective of this project is to develop an automated diagnostic system for bacterial vaginosis with desirable accuracy, similar to that of a human. Most of the components will be written in Python and Lua, and a simple user interface will be developed. A trained classifier will be the core component of this system, with other processing modules supporting the overall flow of the system, including image processing, segmentation, interpretation tools. This system will allow the user to select images and get predictions on the degree of infection.

# 3. Methodology

## 3.1. Overview

In this project, an auxiliary image labelling tool was first produced. Images annotated with positions and types of bacteria, as well as the overall degree of infection by medical experts were then obtained.

| | Score | | |
|---|---|---|---|
| Average Abundance per oil immersion field (1000X magnification) | Lactobacillus morphotypes | Gardnerella and Bacteroides morphotypes | Curved Gram-variable rods |
| 0 | 4 | 0 | 0 |
| < 1 | 3 | 1 | 1 |
| 1 – 4 | 2 | 2 | 1 |
| 5 – 30 | 1 | 3 | 2 |
| > 30 | 0 | 4 | 2 |

Table 1. The Nugent Scoring system [4]

The degree of infection is evaluated according to the Nugent Scoring system [4] (see Table 1), which is based on the average density of three types of bacteria: Lactobacillus morphotypes (scored 0 – 4), Gardnerella and Bacteroides morphotypes (scored 0 – 4), and Curved Gram-variable rods (0 – 2).

| Total Score | Interpretation |
|---|---|
| 0 – 3 | Normal |
| 4 – 6 | Intermediate |
| 7 – 10 | Bacterial vaginosis infection |

Table 2. The interpretation of the Nugent Score [4]

The three scores for each type of bacteria is then summed to a score ranging from 0 to 10, which indicates the overall degree of infection (see Table 2).

After data collection, the development of an automated classification tool using convolutional neural networks typically involves the stages of pre-processing, segmentation, training and evaluation. The images are first pre-processed and segmented [12], and then the segmented areas are then used as training data as well as testing data for the classification task. An evaluation of the performance of the classification task is carried out afterwards. These require both hardware and software support. The aim is to develop a model with high accuracy in estimating the degree of infection by examining a blood film image through different image processing techniques and machine learning models.

## 3.2. Prerequisites

### 3.2.1. Hardware

The access to GPUs (Graphical Processing Units) will be required for training machine learning models. GPUs are optimized for parallel computations, and the nature of machine learning model training, which typically involves a large number of mathematical computations, is highly parallelizable and hence, can be completed in much shorter time using GPUs [13].

### 3.2.2. Software

Machine learning, computer vision and graphics libraries will be required in order to eliminate the time spent in developing such tools and to focus on the development of the diagnostic tool. A number of widely available libraries including Torch, Tensorflow, etc. are identified. Torch [14], a library implemented in the programming language Lua, is currently used in our project because the auxiliary programs, which act as the interface between the machine learning training process and the GPUs available in the Department of Computer Science, are available from previous projects and using them can reduce the risk of incompatibility as well as significantly reduce development time.

## 3.3. Data Collection

The collection of data is essential for the development of a classification system using machine learning methods, where these data are used for both developing and testing the system. Furthermore, the quality of data for training machine learning models has significant effects on the performance of the models [15]. Therefore, in order to achieve satisfactory performance, a tailored dataset is needed for this project. External sources of data tend to be very limited in many ways, including in terms of number, the variation of quality, and the difficulty in fitting the projects' methodology. The collaboration with the Li Ka Shing Faculty of Medicine, the University of Hong Kong, can enable more efficient communication, the production of a tailored dataset for the project, as well as the possibility of obtaining original, unaltered data.

For supervised learning algorithms, data can be separated into two components: the input dataset, which consists of data to be given to the models, and the corresponding output dataset, which consists of facts known about the corresponding input data [16]. In this project, the input dataset consists of a mix of vaginal smear images collected from patients and bacterial colony images, and the output dataset consists of the detailed locations of

bacteria and the degrees of infection for vaginal smear images.

### 3.3.1. Smear Images

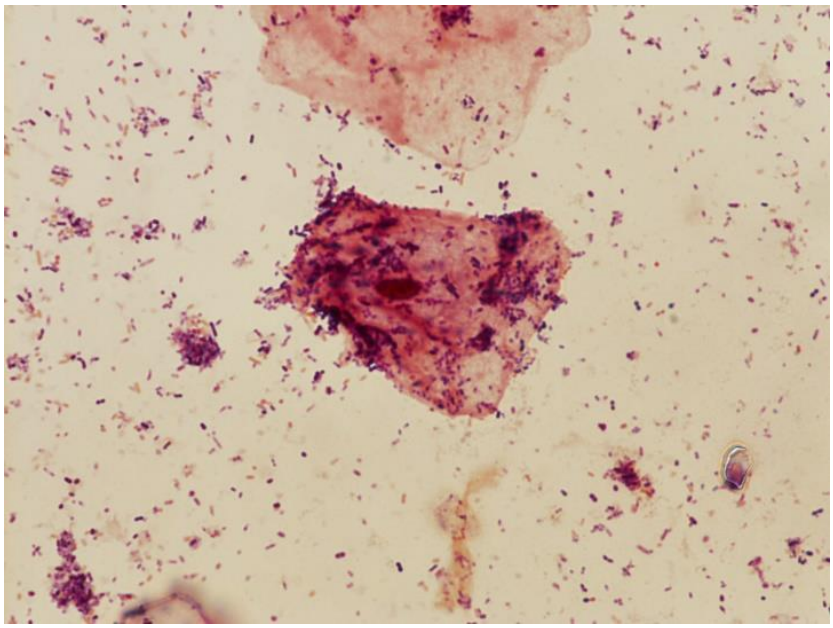

Figure 1. A Gram-stained vaginal smear

Images of Gram-stained vaginal smears (see Figure 1) of varying degrees of bacterial vaginosis infection will be provided by the medical experts from the University of Hong Kong, where the images are collected with the consent from the patients.

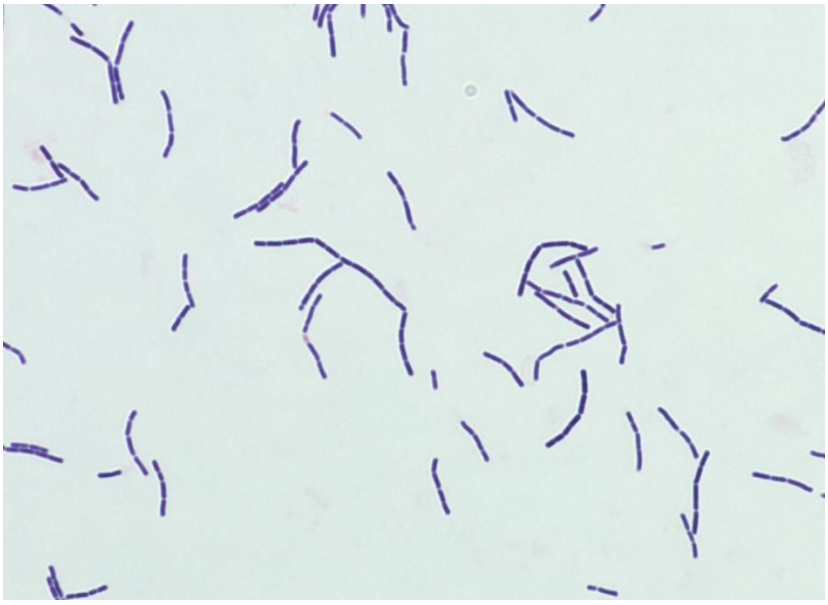

Figure 2. A Gram-stained bacterial colony smear

In addition, images of bacterial colonies (see Figure 2) will also be provided by them. Both types of images are collected to mitigate the shortcomings typically found in each type of images. Vaginal smear images are used in the actual diagnosis of infection by the medical

experts, and hence accurately capture the environment where the bacteria are found in the human body. However, according to the feedback from the microscopists, many of the bacteria present in the smear images cannot be accurately identified by observing the images alone, due to the fact that some of the bacteria may share similar shapes and morphologies at different stages. On the other hand, the bacterial colony images might not accurately capture the morphologies of the bacteria found in human body fluids, the purity of the specimen provides a guarantee on the type of bacteria that is present in these images. Hence, these two types of images are used in conjunction to achieve higher performance.

### 3.3.2. High-level Interpretation of Images

In order to develop an automated diagnostic tool, information about the images, including the degree of infection of the patient, is necessary. The high-level interpretations of the images including the Nugent Score and the overall degree of infection are usually readily available because they are usually recorded during diagnosis.

### 3.3.3. Detailed Labelling of Images and the tool *Clicklable*

However, the detailed labelling of the images, which includes the locations and types of bacteria of individual bacteria is usually not recorded due to the high amount of extra effort required.

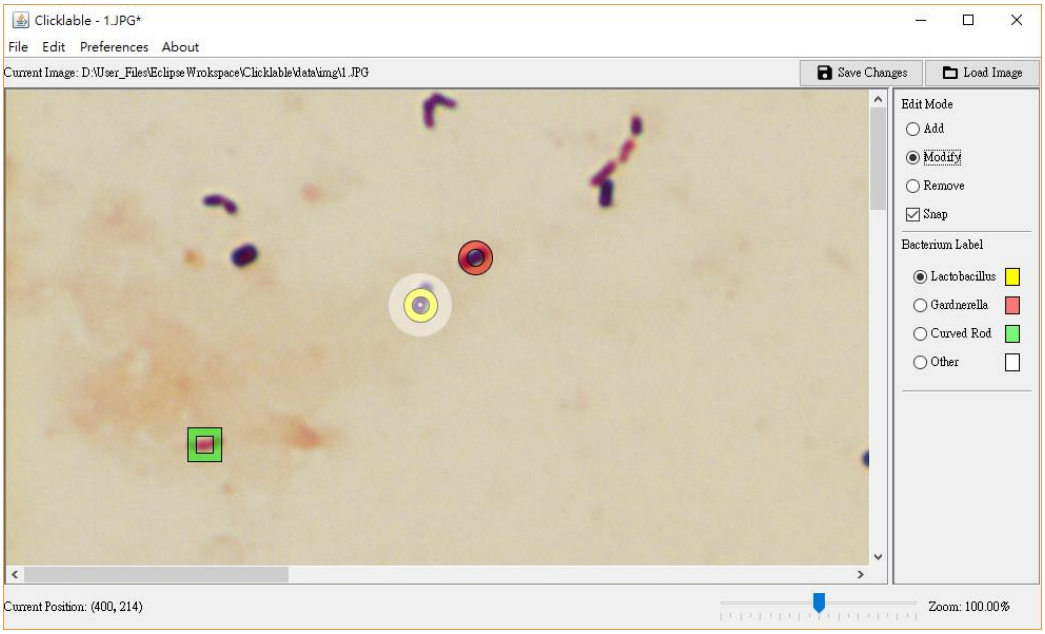


Figure 3. The main user-interface of the data collection tool, *Clicklable*.

Hence, the aforementioned software, *Clicklable* (see Figure 3), was developed to facilitate the annotation.

#### 3.3.3.1. User Interface of *Clicklable*

The tool *Clicklable* allows the medical professionals to load smear images previously captured, and annotate the image by clicking on the locations of individual bacteria. Each annotated location has a shape around it (for example, the red circle which can be seen in the middle of Figure 3).

Furthermore, a number of functionalities are implemented to enhance the user experience. When the user moves the cursor close to an annotation, the corresponding annotation is automatically highlighted (as seen by a semi-transparent white circle around the point, see Figure 4). A popup menu is shown when the user right-clicks on the annotation, and information about the annotation are shown (see Figure 4, where the location and the type of bacteria are shown as the first two items of the menu), together with functionalities to make changes to the label.

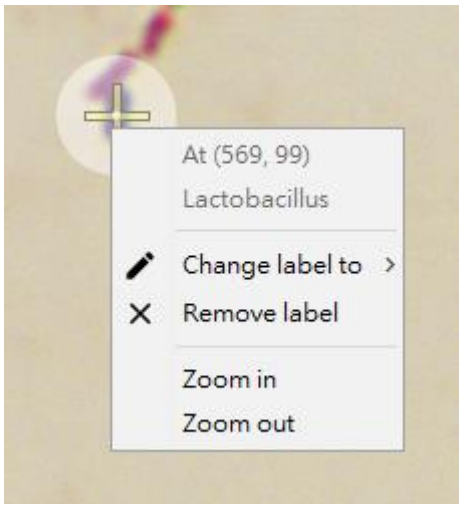


Figure 4. The popup menu in *Clicklable*

In addition, the user-interface elements are highly customisable, where the user can choose the shape, fill and background colours, sizes of the annotations, by changing the corresponding settings (see Figure 5).
These functions allow the user to customise the tool to fit their needs.

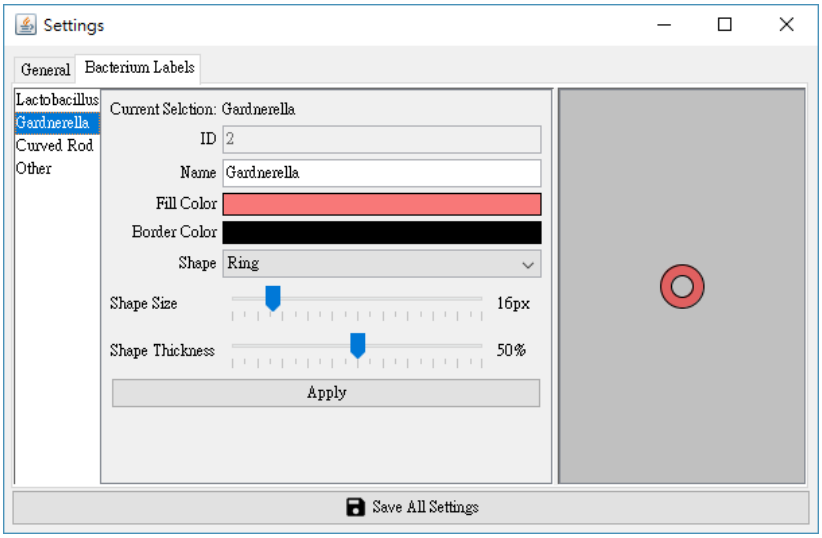


Figure 5. Annotation settings in *Clicklable*

#### 3.3.3.2. Supporting technologies in *Clicklable*

In order to ensure the usability and reliability of the software, various technologies are used in the development of *Clicklable*.
Firstly, Java is the major programming language used in the development of *Clicklable*. Although it requires an installation of the Java Runtime Environment, it is widely available (over 15 billion devices run Java software [17]) and it is independent on the underlying operating system that it is running. This reduces difficulties in distributing this tool to

different platforms and also lowers development time.

Secondly, all the data are stored in a database using a Structured Query Language (SQL) database engine, SQLite. SQLite is a reliable database engine which is resilient against failure and relatively light-weight [18]. Since the data stored in *Clicklable* are simple (only annotation and basic file information), the small amount of extra resources required, and the robustness of the engine are very desirable features.

### 3.4. Image Processing

After the collection of data, different image processing techniques will be employed to reduce the variations between training samples and hence to increase the reliability of the machine learning models. This involves the pre-processing stage and the segmentation stage.

#### 3.4.1. Pre-processing

In this stage, variations between images due to different background lighting, degrees of staining and image acquisition techniques are calibrated.

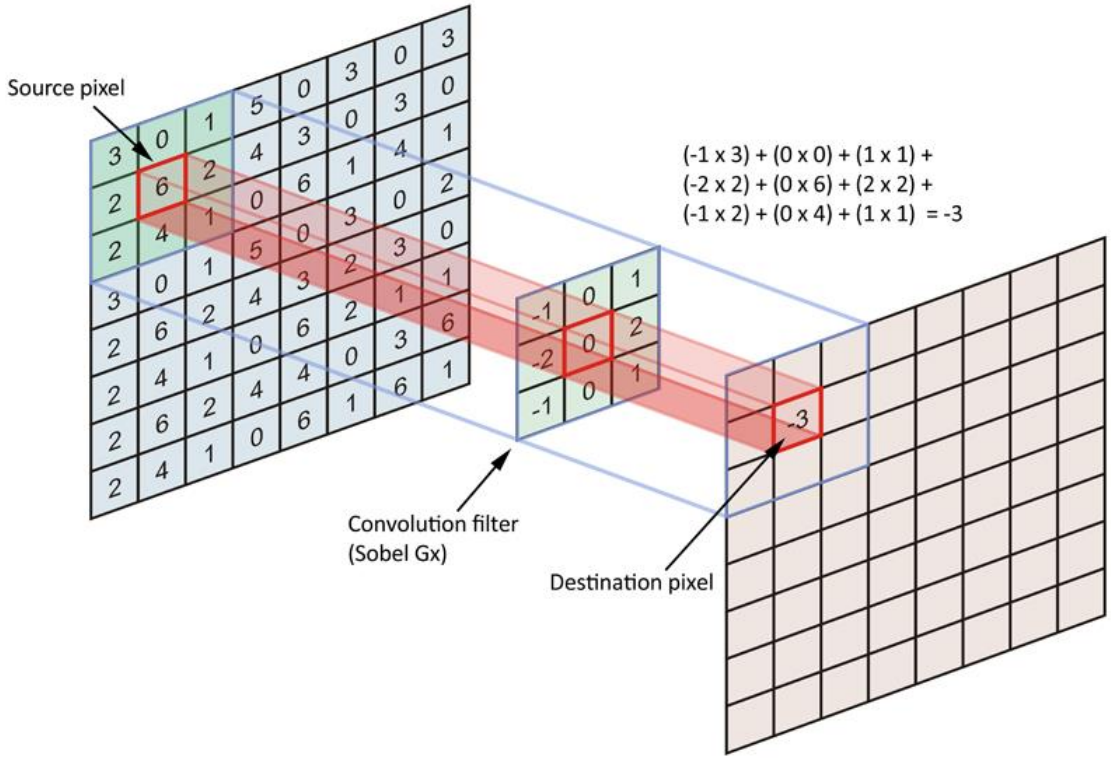


Figure 6. An example of convolution of an image with a convulsion filter [19]

Noises in images can often be effectively removed by applying spatial filtering, a computer vision technique which involves the convolution (or filtering) of the image with a kernel, a weighted matrix (see Figure 6). The convolution of an image *f* with kernel (also called

convolution filter) $g$ is defined on the set of real numbers as [20]:

$$(f * g)[\mathrm{m}, \mathrm{n}] = \sum_{k} \sum_{l} f[\,\mathrm{m} - k, \mathrm{n} - l\,]\,\mathrm{g}\,[k, l]$$

The convolution operation combines the values of the neighbourhood of each pixel in the image. For example, as shown in Figure 6, the highlighted region around the source pixel, is convoluted with the convolution filter to obtain the destination pixel. The calculation is done by multiplying each value in the source region with the corresponding value in the filter, followed by a summation operation (as shown in the top-right corner of Figure 6). This is similar to perceiving an image at a distance, where the information of individual details is not directly perceivable, but rather the general information in an area. The size of the kernel as well as the weights of the kernel are adjusted for different uses. In particular, Gaussian filters are commonly used for reducing noises before edge detection [21].

Level of illumination, on the other hand, can be effectively calibrated by subtracting an empty film (control image) [12], thresholding or analysing the histogram and apply histogram transformation.

Finally, the variations in the scales of images, if not handled properly, could result in meaningless estimations from the model due to that fact that the morphology, in particular, the length of bacteria is particularly important in identifying the identity. This can be solved when the magnification of the microscope is known. However, if the information is not available, it is possible to rely on the assumption that healthy human red blood cells and platelets have similar sizes and a more advanced technique called granulometric analysis [22], can be used to estimate the sizes of the cells and scale accordingly [12]. Different combinations of aforementioned techniques will be applied according to the variations observed generally in the data.

### 3.4.2. Segmentation

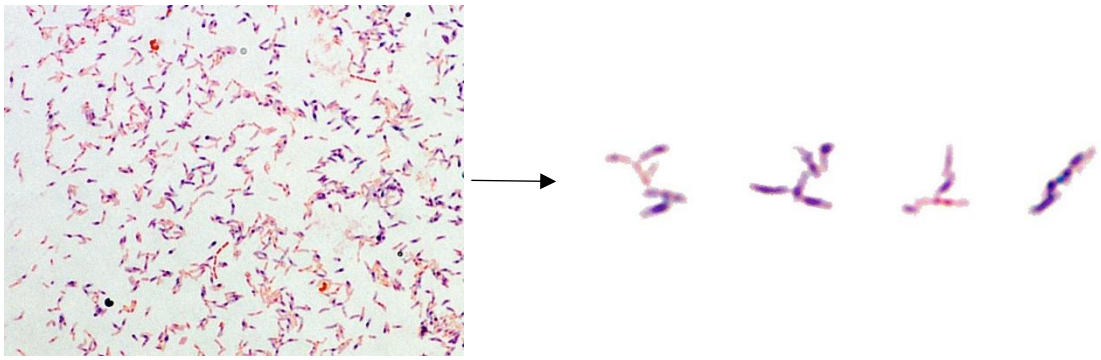

Figure 7. The segmentation of a blood smear image

Segmentation is the process of dividing the image into areas of interest. In this project, the main goal of this stage is to separate the images into small regions which contain one or more bacteria (see Figure 7). Blob detection algorithms, as well as data clustering algorithms, are used for this task in this project. In particular, Maximally stable extremal regions (MSER) and Density-based spatial clustering of applications with noise (DBSCAN) are used.

MSER, proposed by Matas et al. [23], is a blob detection method which was originally proposed for identifying the correspondence areas or objects of images taken from different perspectives. This method is adaptive to a number of common transformations in images taken of the same objects, where the regions identified are invariant to linear transformations of brightness and relatively stable [23]. These are the desired properties and features for the method used for segmenting the blood smear images, such that regions identified are not easily affected by the variation in illumination.

DBSCAN [24], is a data clustering algorithm which is used in identifying clusters in data points such that region of nearby neighbouring data points is identified as a single cluster. This algorithm is robust against outliers, as well as highly flexible in terms of the shapes of the clusters, which are applicable to the blood smear images.

## 3.5. Classification

After the segmentation stage, areas of interests or local frames of the images will be identified. A classifier which distinguishes between the target bacteria, Gardnerella vaginalis, from the rest, or ideally, into separate categories such as lactobacillus, gardnerella, curved rods, etc. will be developed. In addition, different techniques in determining the number of bacteria in each area of interest will also be explored. A number of machine learning algorithms including neural networks, support vector machine and fuzzy logic are potential candidates for developing the classifier. In this project, convolutional neural networks will be

used as the main technique.

### 3.5.1. Artificial Neural networks

Artificial neural network [25] is a machine learning algorithm, where a neural network is formed by combining a collection of artificial neurons. These artificial neurons are modelled by a mathematical function (see Figure 8), defined on the set of real numbers from inputs $x_0, x_1, \dots, x_i$, weights $w_0, w_1, \dots, w_i$, bias $b$, and activation function $f$ to output $y$, as [26]:

$$y = f\left(\sum_i w_i x_i + b\right)$$

This aims to model a neuron in the human brain (see Figure 9), the fundamental unit of computation of the human brain.

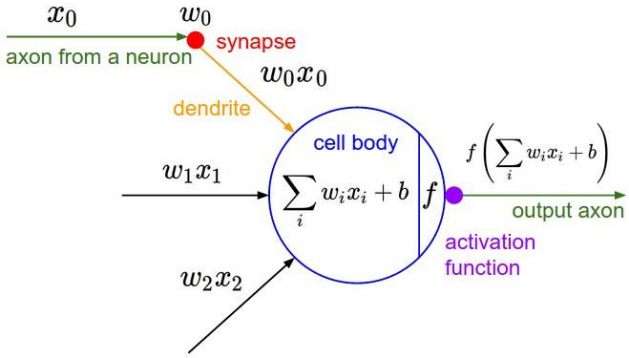


Figure 8. The mathematical model of a neuron (a node in artificial neural networks) [21]

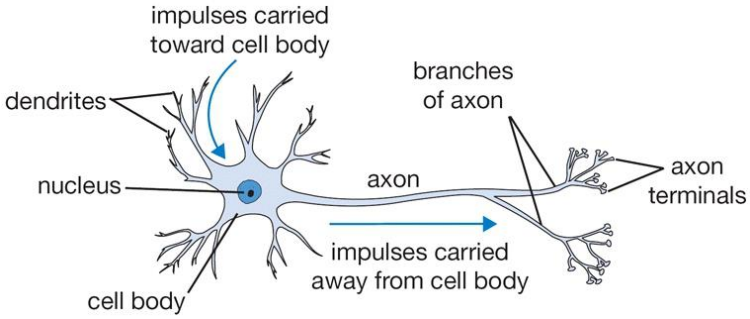


Figure 9. The diagram of a human neuron [21]

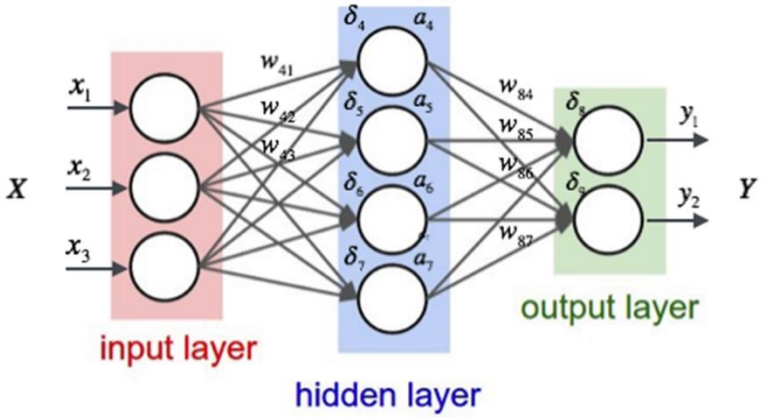


Figure 10. A 3-layer neural network [21]

These artificial neurons are then connected, to form an artificial neural network. They are separated into three groups: input, output and hidden (see Figure 10). For example, in Figure 10, there are 3 neurons in the input layer, 4 neurons in the hidden layer and 3 neurons in the output layer. These nodes are interconnected such that values except for the input nodes are calculated based on the values of other nodes and mutable parameters. Supervised learning

for classification, where each learning sample is provided with its desired output, involves finding the parameters in the network such that when presented with new data, the network is able to generate desired output with high accuracy. This generalization process typically does not require handcrafted features or weightings of the input, where the network "learns" by inferring the relationship between the inputs and the outputs from the training samples. This significantly reduces the necessity of expertise in the related area for tailoring the important features. A variety of different architectures have been proposed, mainly differing in how the network is structured, how the parameters are tuned, and what functions are used in the calculations. Different architectures are used based on the purpose of such network.

### 3.5.2. Convolutional neural networks

Convolutional neural networks [27], a type of artificial neural networks, make use of the convolution operation (as presented earlier in section 3.4.1.) in addition to standard linear operations.

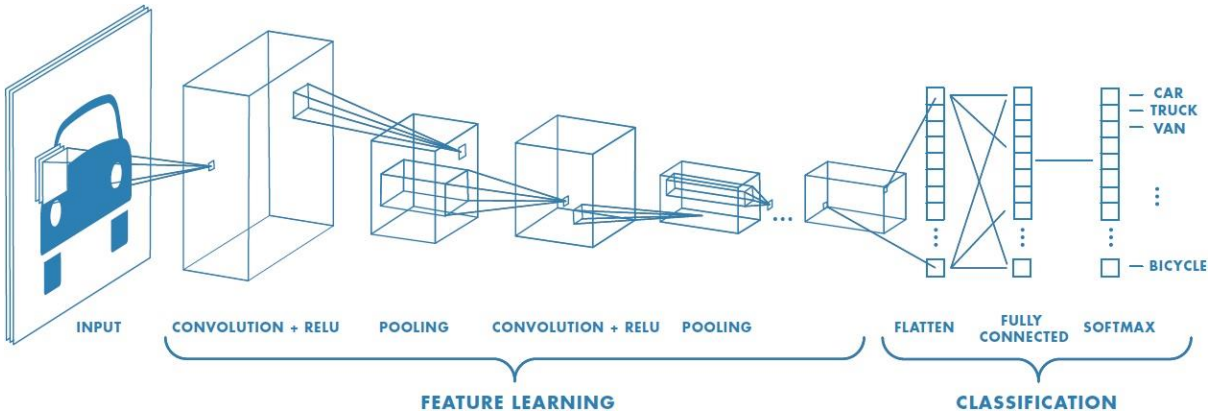


Figure 11. Example of a network with many convolutional layers. [28]

A convolutional neural network for classification typically starts with the input image as the input nodes and subsequently applies convolutions (filtering) and sub-sampling (max pooling) on the values until the output layer which indicates the likelihood of being in a certain category is reached (see Figure 11). During training, the weightings (parameters) of the nodes are adjusted to fit the expected outcome. This type of architecture is very effective in dealing with image inputs, due to the nature of the convolution operation which uses values from local neighbourhoods of the image. The first pattern recognizer which achieved human-level performance on several tasks was based on this learning method.

Since the segmented areas are essentially a part of the image and because of the successes seen in other similar projects, convolutional neural network is a strong candidate for the architecture of the classifier in the project.

However, A range of hyper-parameters of such model including the number of layers, size of filters, loss function, etc. will need to be tuned and compared.

### 3.6. Interpretation

After identifying the number of different bacteria present in the blood smears, a final data analysis which estimates the degree of infection will be done. The number of bacteria in each category of the Nugent Score System [4] will be identified and an overall interpretation based on the same system will be made.

### 3.7. Performance Evaluation

The performance of the models will be evaluated by determining the accuracy of the predictions for images which are not used in training process. Performance evaluation can be separated into two stages: validation and testing. Data collected are typically separated into three sets accordingly: the training set, the validation set, and the testing set, where only the training set is used for training the model. The validation set is used to evaluate the performance of the model, and the hyper-parameters including the architecture, number of learning iterations, are tuned to maximize the performance. The testing set, on the other hand, is reserved for the final evaluation after the hyper-parameters are tuned and is used to reflect the generalizability of the model. The reason for separating the performance evaluation into two stages is that tuning hyper-parameters to maximize performance actually leaks certain information about the testing data into the model, which could lead to a false performance of the model because the model might only show good performance on the current dataset and fail to generalize [29]. Hence, an exclusive set of testing data is reserved for the final evaluation of the model.

In terms of measurements, the accuracy of the classifier for classifying individual segmented image, a coarse accuracy of high-level interpretation which is determined by correctly identifying the general degree of infection (Normal, Intermediate and Infected) of each image as a whole, and a fine accuracy of high-level interpretation which is determined by accurately predicting the Nugent Score (from 0 to 10) [4] of each image as a whole will be used as the performance metrics.

# 4. Results

## 4.1. Work completed

### 4.1.1. Data collection

The first batch of images was obtained during meetings held in March and April 2017. This includes 40 bacterial colony images and 31 vaginal smear images. The smear images are annotated with high-level interpretations, including Nugent Score and overall degree of infection only.

The second batch of images was obtained during meetings held in October 2017. This batch consists of 119 vaginal smear images, all annotated with high-level interpretation. After the development of the data labelling tool, the collection of detailed labelling started in late October 2017, with currently 30 images annotated in a high level of detail.

### 4.1.2. Data Labelling Tool

The development of the data labelling tool, *Clicklable*, was completed and distributed to the medical experts in October 2017. The feedback from the medical professionals was satisfactory, and the user interface will be refined according to future requirement and feedback from them.

### 4.1.3. Diagnostic system

The first phase of the project, which involved the development of the first prototype of the diagnostic system was completed in September 2017. Basic image segmentation tools were then developed, and the auxiliary programs for training the model learning models were also adapted for the project. The first classifier prototype for the type of bacteria was trained and evaluated.

The second phase of the project, which includes the review of the scope of the project, exploring ways to improve the performances of the diagnostic tool, and streamlining the auxiliary tools are currently under development.

## 4.2. Performance of the first diagnostic tool prototype

The first batch of data collected consists of 71 images, and out of all images, 31 of those, which were collected from the patients, the Nugent Score is also available for them and hence are used for testing. All images were directly segmented using MSER. The segmented images which originated from the bacterial colonies were used as the training and evaluation data for

the machine learning model, with 30% of the data reserved for validation. All the remaining segmented images, which were originally collected from the patients, were then used as testing data. These images were not used for training nor validation due to the lack of detailed labelling of individual bacteria in the patient images. The first classifier prototype was then trained and fine-tuned according to the performance on the validation set.

| | | Predicted type of bacteria by the model | | | |
|---|---|---|---|---|---|
| | | Lactobacilli | Gardnerella | Curved rods | Other |
| Actual type of bacteria | Lactobacilli | **499** | 6 | 1 | 2 |
| | Gardnerella | 0 | **559** | 3 | 4 |
| | Curved rods | 0 | 2 | **589** | 0 |
| | Other | 2 | 5 | 0 | **63** |

Table 3. Validation results on type of bacteria in segmented images of the first classifier prototype (Accuracy = 1710 / 1735 = 98.6%)

The first classifier prototype has an accuracy of 98.6% in the validation stage, with the detailed performance shown in Table 3. Out of the 1735 segmented regions within the validation set, 1710 were accurately classified by the model. Some confusion between bacteria types is observed, for example, there are 6 segmented images of bacteria type Lactobacilli wrongly classified as Gardnerella by the model (as shown in the cell by the column “Gardnerella” and row “Lactobacilli” in Table 3), but overall this is a highly satisfactory performance, which indicates the hyper-parameters of the model is well tuned.

| | | Estimated degree of infection by the model | | |
|---|---|---|---|---|
| | | Normal | Intermediate | BV Infection |
| Actual degree of infection | Normal | **8** | 0 | 0 |
| | Intermediate | 5 | **3** | 0 |
| | BV Infection | 6 | 6 | **3** |

Table 4. Testing results on degree of infection of the first classifier prototype (Accuracy = 14 / 31 = 45.1%)

However, in the final evaluation using patient images, the model only achieved an exact-match accuracy of 45.1% (see Table 4) for the coarse high-level interpretation. Out of the 31 patient images, only 14 of them are accurately estimated the degree of infection. For the remaining 17 patient images, all of them are underestimated by the model.

| | | Estimated Nugent Score by the model | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Actual Nugent Score | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | **2** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | 0 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 3 | 2 | 1 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 4 | 0 | 1 | 1 | 1 | **0** | 0 | 0 | 0 | 0 | 0 | 0 |
| | 5 | 0 | 1 | 1 | 1 | 0 | **0** | 1 | 0 | 0 | 0 | 0 |
| | 6 | 0 | 0 | 0 | 0 | 0 | 1 | **0** | 0 | 0 | 0 | 0 |
| | 7 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | **0** | 0 | 0 | 0 |
| | 8 | 0 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | **1** | 0 | 0 |
| | 9 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | **0** | 0 |
| | 10 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | **0** |

Table 5. Testing results on Nugent Score of the first classifier prototype (Accuracy = 3 / 31 = 9.6%)

Furthermore, the model only achieved an exact-match accuracy of 9.6% (see Table 5) for the fine high-level interpretation. Almost all the estimations by the model are underestimations, which can be seen from the majority of values lying below the diagonal. It can be inferred that the model tends to underestimate the severity of the infection.

This drastic difference in performance indicates that the model does not generalize well, and may have the problem of overfitting.

# 5. Limitations and Future Direction

## 5.1. Problems identified

The significant difference between validation and testing performances of the first diagnostic tool prototype indicates potential flaws in the development of the tool, especially in terms of the use of data.

### 5.1.1. Limited amount of data

The availability of images is relatively low in this project, where the expected number of images is below 300. This restriction on data makes it impractical to directly develop a classifier using images as a whole.

### 5.1.2. Similarity among bacteria

Figure 12A. Segmented area containing a bacterium type “Lactobacillus”

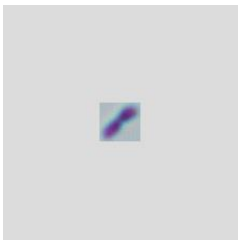

Figure 12B. Segmented area containing a bacterium type “Gardnerella”

Under certain conditions, it is difficult to distinguish between the bacteria types Gardnerella and Lactobacillus (see Figure 12A & 12B) due to the similarity of morphology and variations in staining. The medical experts who provide us with the data also confirm the ambiguity. This could potentially undermine the feasibility of developing a highly accurate automated diagnostic system.

### 5.1.3. High variation in smear images

As the smear images are obtained in batches and are prepared by humans, there are sometimes significant variations in the degree of staining and level of illumination.

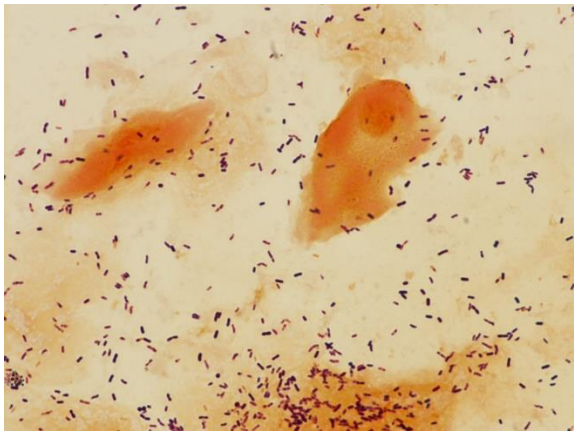

Figure 13A. A patient smear image in image batch 1

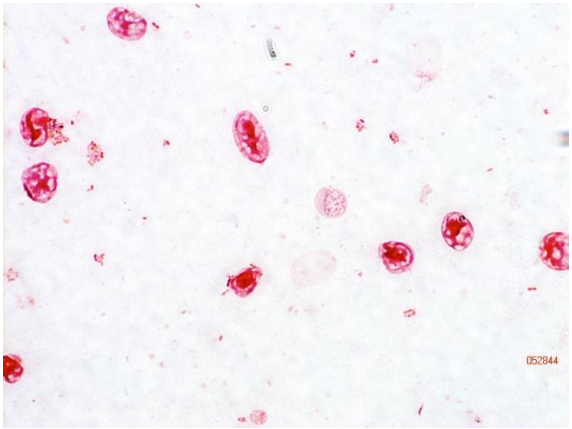


Figure 13B. A patient smear image in image batch 2

Among some of the images, the variation is very noticeable (see Figure 13A & 13B), where the levels of staining are significantly different. This level of variation might not be properly handled by the image processing algorithms in the first prototype.

#### 5.1.4. Discrepancies in morphology

A noticeable difference in morphology of the same type of bacteria in different environments (colony and patient) is identified. Due to the fact that the model is trained using only the colony images, this difference could significantly undermine the performance of the model, and this could be the explanation for the difference in performances.

## 5.2. Future Direction

In order to mitigate the problems encountered, changes to the development process of the diagnostic tool are made in the second phase of the project.

First of all, a larger, more detailly labelled set of images is sought from the medical experts.

In the development of the second classifier prototype, both colony and patient images will be first pre-processed to reduce the possible variations in the degree of staining and illumination, and then segmented for both training and testing purposes, instead of using only colony images for training. Also, all three measures of performance will be used in both validation and testing to better reflect performance.

Furthermore, different techniques in image processing will be employed to better reduce the variation between images, especially in the degree of staining and level of illumination.

In addition, the set of tools used in different steps of development (pre-processing, segmentation, training, evaluation) will be streamlined such that it allows faster development cycles as well as easier integration for the final system.

Moreover, after the development of a diagnostic system with satisfactory performance, the limits of the development of such tools as well as general strategies in applying similar methodologies in the medical contexts will be explored, in the third phase of the project.

## 5.3. Project Schedule

This project is separated into three phases, each with important deliverables and milestones at completion.

#### 5.3.1. Schedule of Phase 1

| Item | Finished on |
|---|---|
| Training of the first batch of machine learning models and the analysis of their performance | 29 September 2017 |
| Submission of detailed project plan and construction of project web page | 1 October 2017 |
| Meetings with medical experts to get feedback for the performances of the models | 14 October 2017 |

#### 5.3.2. Schedule of Phase 2

| Item | Finish by |
|---|---|
| Further investigation into image processing modules | 21 October 2017 |
| Collection of new data from the medical experts | 14 November 2017 |
| Training of the second batch of machine learning models based on new data and new image processing techniques | 25 January 2018 |
| First presentation | 29 January 2018 |

#### 5.3.3. Schedule of Phase 3

| Item | Finish by |
|---|---|
| Exploration of the limits of machine learning techniques in medical contexts | 14 March 2018 |
| Propose potential improvements and general strategies in applying such techniques | 31 March 2018 |
| Final fine-tuning of the integrated diagnosis system | 31 March 2018 |
| Project exhibition | 2 May 2018 |

# 6. Conclusion

Diagnostic microscopy is the gold standard for many infections, including bacterial vaginosis. However, the high costs and the dependence on human expertise in microscopic diagnosis are still present. There have been successful cases of developing an automated diagnostic tool for some common infections using machine learning or related techniques. This shows a promising possibility for the success in developing one for bacterial vaginosis. This project explores the practicability of using machine learning and computer vision techniques in the development of such tool. In order to facilitate the development of similar tools in medical contexts, this project will also act as an example for demonstrating the limitations as well as general strategies in the development process.

In the first phase of this project, the first diagnostic tool prototype was developed using convolutional neural networks and trained using colony images. It has contrasting performances in validation (accuracy of 98.6%) and testing (accuracy of 45.1% for the degree of infection, 9.6% for Nugent Score), suggesting the possibility of overfitting in the prototype. These results show that the variations of the image acquisition process, the variations in the morphology of bacteria in different environments might not be carefully handled, and the labelling of data might not be detailed enough to successfully develop an automated diagnosis tool. Changes in the development process are made in the second phase of this project accordingly. These development processes will be analysed and used as examples in the last part of this project.

Due to the restriction on the amount of data, this project does not tackle the classification task directly using whole blood smear images. With a higher availability of data, the approach of using blood smear images as a whole might be used in future research, and the difference between the performances and complexities could be explored. Also, further investigation into the morphologies of bacteria in different environments can be done to better understand the features of target bacteria. More tailored algorithms for different parts of the system can then be used, which might further improve the performance.

# 7. References

[1] C. Kenyon, R. Colebunders and T. Crucitti, “The global epidemiology of bacterial vaginosis: a systematic review,” *American Journal of Obstetrics and Gynecology,* vol. 209, no. 6, pp. 505-523, 2013.

[2] E. H. Koumans, M. . Sternberg, C. . Bruce, G. M. McQuillan, J. S. Kendrick, M. Y. Sutton and L. E. Markowitz, “The prevalence of bacterial vaginosis in the United States, 2001-2004; associations with symptoms, sexual behaviors, and reproductive health,” *Sexually Transmitted Diseases,* vol. 34, no. 11, pp. 864-869, 2007.

[3] J. Atashili, C. Poole, P. Ndumbe, A. Adimora and J. Smith, “Bacterial vaginosis and HIV acquisition: a meta-analysis of published studies,” *AIDS,* vol. 22, no. 12, pp. 1493-1501, 2008.

[4] R. P. Nugent, M. A. Krohn and S. L. Hillier, “Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation,” *Journal of clinical microbiology,* vol. 29, no. 2, pp. 297-301, 1991.

[5] B. Sha, H. Chen, Q. Wang, M. Zariffard, M. Cohen and G. Spear, “Utility of Amsel Criteria, Nugent Score, and Quantitative PCR for Gardnerella vaginalis, Mycoplasma hominis, and Lactobacillus spp. for Diagnosis of Bacterial Vaginosis in Human Immunodeficiency Virus-Infected Women,” *Journal of Clinical Microbiology,* vol. 43, no. 9, pp. 4607-4612, 2005.

[6] R. Chawla, P. Bhalla, S. Chadha, S. Grover and S. Garg, “Comparison of Hay’s Criteria with Nugent’s Scoring System for Diagnosis of Bacterial Vaginosis,” *BioMed Research International,* vol. 2013, pp. 1-5, 2013.

[7] World Health Organization., “Microscopy,” 14 March 2017. [Online]. Available: http://www.who.int/malaria/areas/diagnosis/microscopy/en/. [Accessed 27 September 2017].

[8] K. C. Hazen., “Microscopy,” October 2016. [Online]. Available: http://www.merckmanuals.com/professional/infectious-diseases/laboratory-diagnosis-of-infectious-disease/microscopy. [Accessed 13 February 2017].

[9] J. Bennett, Mandell, Douglas, and Bennett's principles and practice of infectious diseases, Philadelphia: Elsevier/Saunders, 2015.

[10] J. A. Quinn, R. Nakasi, P. K. Mugagga, P. Byanyima, W. Lubega and A. Andama, "Deep convolutional neural networks for microscopy-based point of care diagnostics," in *Machine Learning for Healthcare Conference,* 2016.

[11] O. Kraus, J. Ba and B. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics,* vol. 32, no. 12, pp. i52-i59, 2016.

[12] F. Tek, A. Dempster and I. Kale, "Computer vision for microscopy diagnosis of malaria," *Malaria Journal,* vol. 8, no. 1, p. 153, 2009.

[13] D. Steinkraus, I. Buck and P. Simard, "Using GPUs for machine learning algorithms," in *Eighth International Conference on Document Analysis and Recognition,* 2005.

[14] R. Collobert, C. K. K. Farbet and S. Chintala, "torch," [Online]. Available: http://torch.ch/. [Accessed 22 10 2017].

[15] V. Sessions and M. Valtorta, "The Effects of Data Quality On Machine Leraning Algorithms," in *Conference: Proceedings of the 11th International Conference on Information Quality,* Cambridge, MA, USA, 2006.

[16] MathWorks, "Introducing Machine Learning," 2016. [Online]. Available: https://www.mathworks.com/content/dam/mathworks/tag-team/Objects/i/88174_92991v00_machine_learning_section1_ebook.pdf. [Accessed 30 November 2017].

[17] Oracle, "Go Java," 29 August 2016. [Online]. Available: https://go.java/index.html. [Accessed 30 November 2017].

[18] SQLite, "SQLite Home Page," 2017. [Online]. Available: https://sqlite.org/index.html. [Accessed 30 November 2017].

[19] A. Durville, "Computer Vision with Convolution Networks," 11 February 2017. [Online]. Available: https://github.com/OKStateACM/AI_Workshop/wiki/Computer-Vision-with-Convolution-Networks. [Accessed 23 October 2017].

[20] S. Damelin and W. Miller, The Mathematics of Signal Processing, Cambridge University Press, 2011.

[21] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 6, pp. 679-698, 1986.

[22] P. Maragos, "Pattern spectrum and multiscale shape representation," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 11, no. 7, pp. 701-16,

1989.

[23] J. Matas, O. Chum, M. Urban and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions.,” *Image and vision computing,* vol. 22, no. 10, pp. 761-767, 2004.

[24] M. Ester, H. P. Kriegel, J. Sander and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” *Kdd,* vol. 96, no. 34, pp. 226-231, 1996.

[25] B. Yegnanarayana, Artificial neural networks, PHI Learning Pvt. Ltd., 2009.

[26] Stanford University, “CS231n Convolutional Neural Networks for Visual Recognition,” 28 November 2017. [Online]. Available: http://cs231n.github.io/neural-networks-1/. [Accessed 30 November 2017].

[27] L. Yann, “LeNet-5, convolutional neural networks,” [Online]. Available: http://yann.lecun.com/exdb/lenet/. [Accessed 21 October 2017].

[28] MathWorks, “Convolutional Neural Network,” 2017. [Online]. Available: https://www.mathworks.com/discovery/convolutional-neural-network.html. [Accessed 23 October 2017].

[29] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *Ijcai,* vol. 14, no. 2., pp. 1137-1145, 1995.