October University for Modern Sciences and Art

Faculty of Computer Science

Graduation Project

Title: AMOI (Ai-aided Microorganisms Identification)

Supervisor: Dr. Ehab Emam Zakaria

Name: Omar Ynis

ID: 191433

Abstract

Microorganisms are microscopic organisms that can only be observed under a microscope. Some microbes are useful to humans and are employed in a variety of fields. This research aims to create a system that will aid in the identification and differentiation of these bacteria from a variety of species. Microorganism identification is one of the many procedures that scientists and researchers go through in their daily lives. Unfortunately, they either employ heavy machinery and pricey software, or they use the old school method of manually identifying microorganisms. This research aims to create an automatic system capable of identifying microorganisms and making it available easier without heavy machines, while maintaining the speed and precision. Many approaches and techniques will be used in order find the best accuracy and at the same time make it possible for everyone to use the system. Image processing techniques like edge detection, segmentation, color detection and many others will be used to preprocess the data before applying feature selection to decrease the load of the data which will be further processed in a CNN architecture to return an acceptable precision output.

Table of Contents

Table of Figures

Table of Tables

Chapter 1:

Introduction

1.1 Introduction

Microorganisms are small organisms that can only be seen with the help of a microscope. The planet is inhabited by trillions of microorganisms i.e., viruses, fungi, algae, bacteria, etc. which are a key factor not only for the human body, but also the environment itself. The primary inhabitants and designers of our entire biosphere are bacteria. Furthermore, they are the oldest and most prevalent life forms on the planet. (A. James, D. Dennis et al., 2007). Only a small percentage of microorganisms are beneficial to humans, and are used for multiple purposes like food fermentation, food processing, medical field, agriculture, and many other fields. Moreover, these organisms are grown into plates called *Petri Dish* in order to identify and research them better. Technology is slowly changing people's lives all across the world, with healthcare playing a major role in these changes. Hence why, the ability to share health information with accurate diagnosis and the reduction of mistakes and errors found in data through manual records became a major focus. (G. Chadha, A. Srivastava et al., 2020). This project will consist of a system that helps identify and distinguish these microorganisms from many other different species. In order to achieve such results, a variety of strategies and approaches of deep learning and image processing will be applied on digital images. According to different research, scientists and researchers go through a series of processes on their everyday life, and microorganisms identification is one of them. Unfortunately, they either make use of bulky machines and expensive software or go the old school style and identify microorganisms manually.

1.2 Problem statement

As mentioned previously, microorganisms are used in many fields, and considering the vast variety of species still unknown to humankind, further research and understanding of microorganisms and their identification is a necessity. Moreover, A. James, D. Dennis et al., also stated that there is a need for investments in basic research on bacteria, and individual research groups must continue to play a significant role in trying to discover new information about these interesting creatures. Hence why because of its high demand, a faster and easier method to make it possible for anyone to study them is required.

1.3 Objective

This project objective is to provide a system that processes digital images of microorganisms without the use of expensive machines and differentiate among the vast species while still making the whole process automatic and easy-to-use for everyone. In addition, this project also aims to achieve a lightweight system that runs smoothly without depending heavily on resources.

1.4 Motivation

The application of computer vision in the medical field is an interesting way of helping humankind with the help of computers, which in the current period is ideal since technology is advancing day by day. Moreover, making the identification of microorganisms available and easier for everyone, will provide more opportunities for the people who are in interested in this topic. In addition, this will encourage future research and provide many helpful discoveries in the future.

1.5 Thesis layout

In this thesis, the first chapter will provide an introduction about the project and its aim. Then, the second chapter will provide a literature review and a background of the previous work in the same area of research. At last, the third chapter will discuss the materials/tools, methods and environments used, and system architecture overview for the project.

Chapter 2:

Background and Literature Review

2.1 Background

Computer vision is an artificial intelligence field and subset of machine learning that allows computers and systems to extract useful information from digital images, videos, and other visual inputs, as well as to conduct actions or make suggestions based on that information. Nowadays is vastly used in many fields like mobile applications, medicine, and many other fields, but the most important factor is identification of objects. Computer Vision works similar to the human vision, in order to recognize objects by learning from experience. Over the years many approaches were made, and many machine learning techniques were used to improve the precision of the identification.

2.1.1 SVM Algorithm

According to K. Vojislav (2005), SVM (Support Vector Machine) is a supervised machine learning approach that may be used to address issues like classification and regression. It is, however, mostly used to tackle categorization problems. Each data item is represented as a point in n-dimensional space (where n is the number of features), with the value of each feature being the SVM algorithm's value for a certain position. Then we locate the hyper-plane that clearly separates the two classes to complete categorization.

2.1.2 Artificial Neural Networks

Artificial neural networks (ANNs) are artificial dynamic systems that are highly influenced with how the human brain works. They have the capacity to alter their internal organization in response to a function objective. Moreover, due to their ability to recreate the fuzzy rules governing the ideal solution for problems of this type, they are particularly well suited for dealing with nonlinear difficulties. (G. Enzo & B. Massimo, 2008).

2.1.3 Convolutional Neural Networks

As stated by C. Chia-Yen, & B. Chloe et al. (2020), CNNs (Convolutional Neural Networks) are a type of deep neural networks that have shown to be particularly effective at classifying images. It uses the Convolution technique, which is a mathematical operation on two functions that results in a third function that indicates how the shape of one is impacted by the shape of the other.

2.2 Previous Work

2.2.1 Related Work 1

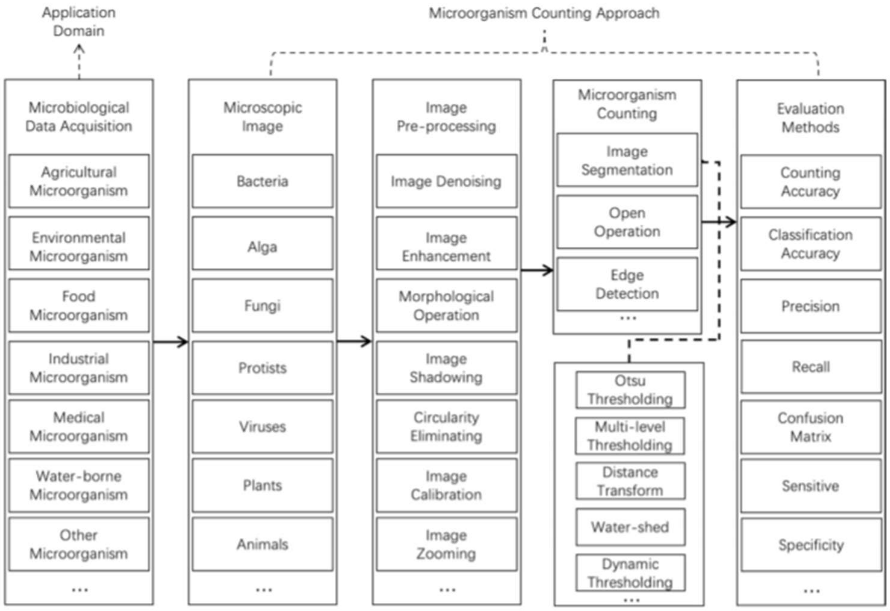

Figure 1: Architecture of Related Work #1

Source (https://doi.org/10.1007/s10462-021-10082-4)

Zhang, J., Li, C., Rahaman, M.M. et al., (2021), conducted a study to determine the best approach for counting microorganism. This study used retrieves digital images of different types of microorganisms with the purpose of counting the entities inside these images. As shown by Figure 1, this study certain approaches to tackle this problem. The approach it uses consists of five steps: image pre-processing, microscopic image, microorganism counting methods, evaluation methods, microbiological data acquisition.

This research retrieved their dataset from different related works made during the 1980s. It consisted of digital images of different microorganisms.

This data is further processed and sent to the ANN in order to identify and count the microorganisms. According to the type of microorganism, it uses different image processing enhancing techniques i.e., edge detection, color segmentation, thresholding, watershed, etc.

At last, this research uses accuracy as its evaluation factor. As a result, it was concluded that when applying ANN to their model, it produces accuracy as high as 95%.

2.2.2 Related Work 2

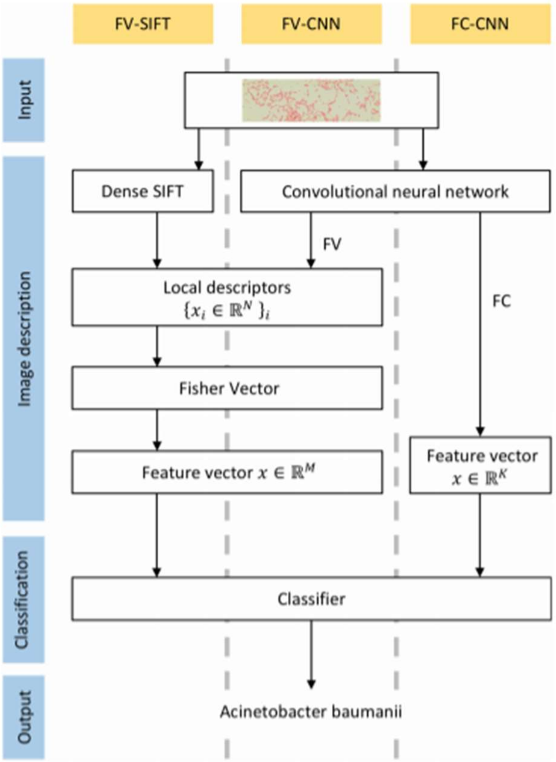

Figure 2: Architecture of Related Work #2

Source (https://doi.org/10.1371/journal.pone.0184554)

Another research was also conducted by B. Zieliński, A. Plichta (2017). Figure 2 shows how this research uses a combination of SIFT and CNN for feature extraction, which is further converted into fisher vectors. Moreover, these fisher vectors are used in a classifier for correctly identify bacteria. The exact processes of this approach are the extraction of image descriptors, single feature extraction, and classification using Support Vector Machine, also known as SVM.

This research retrieved the necessary input data from a public dataset which is composed of digital images of bacteria taken under the microscope. In addition, the bacteria retrieved were experimented with a GRAM test beforehand in order to divide them into two different categories, Gram negative and Gram positive. Moreover, the different categories can be distinguished by the color purple for Gram negative and by the color red for Gram positive. Furthermore, 11 types of bacteria and their subspecies were used in the dataset. http://misztal.edu.pl/software/databases/dibas/

This research makes use of feature selection in order to preprocess the dataset before being classified to remove frivolous or redundant information, which will decrease the size of the dataset and improve the prediction accuracy.

The model which uses a concatenation of Fisher Vector and CNN shows the highest accuracy resulting in approximately more than 90% precision.

2.2.3 Related Work 3

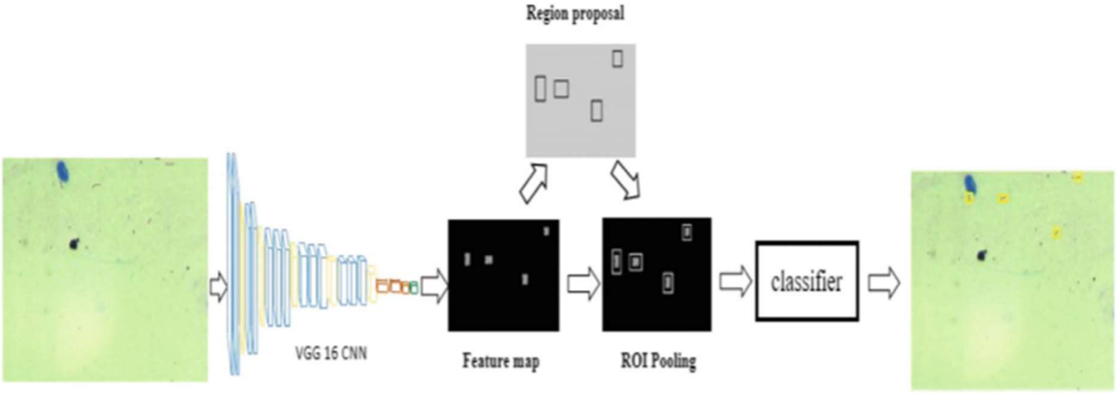

Figure 3: Architecture of Related Work #3

Source (https://doi.org/10.1007/978-3-030-31332-6_24)

Another work was conducted by M. Melegy, D. Mohamed et al., (2019). It can be seen from Figure 3, how this work approached the solution of their problem. This research consists of localize and identify Tuberculosis Bacilli (TB) by the use of Faster Region-based Convolutional Neural Network framework, also known as RCNN. In addition, the architecture used for this research is acquiring digital images of TB from under the microscope and use them as an input to VGG 16 CNN. After applying the RCNN to this model, this architecture will be able to correctly spot and classify TB.

This research retrieved digital images of Tuberculosis Bacilli taken under both conventional microscope and a fluorescence microscope. The TB were stained with a method called Ziehl-Neelsen (ZN) in order to make the TB recognizable for detection. This method is used to differentiate between acid-fast and non-acid-fast bacteria. Moreover, ZN staining works by staining non-acid-fast bacteria and turn them blue, while acid-fast bacteria retain their red color thus helping in identify acid-fast bacteria such as TB. The dataset used was ZNSM-iDB [11].

Furthermore, for evaluating their approach, this research compared their work with other methods by using Recall, Precision, and F-score as factors. Consequently, the use of R-CNN resulted in 98.3% recall, 82.6% precision, and 89.7% f-score. By the provided comparison, the R-CNN proved to exceed the other methods used in every factor.

2.2.3 Related Work 4

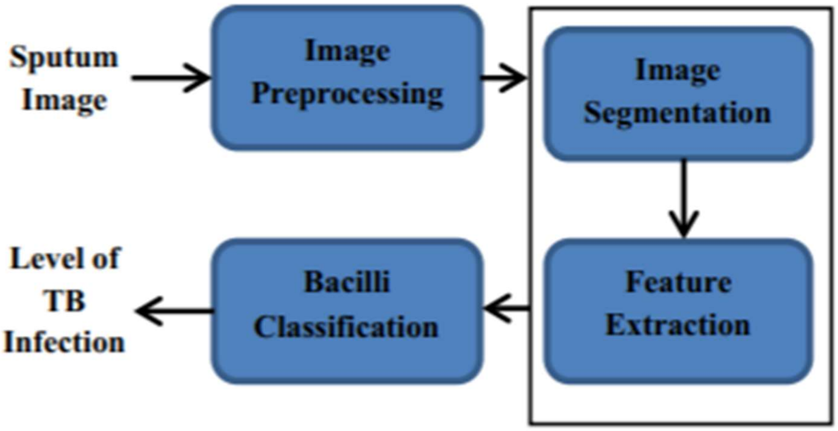

Figure 4: Architecture of Related Work #4

Source (https://doi.org/10.1007/s11760-019-01509-1)

Lastly, another research also tackled the TB identification problem, which was conducted by K. S. Mithra & W. R. Sam. As shown by Figure 4, this research uses an approach that consists of preprocessing sputum images and the apply feature extraction after implementing image segmentation. Consequently, this approach leads to the classification of bacilli. The image processing techniques used were noise reduction and intensity modification. Furthermore, Channel Area Thresholding (CAD) was proposed for image segmentation which is then followed by SURF for feature selection. Moreover, after the feature extraction is done the data is then used as an input for RBM, also known as Restricted Boltzmann Machines.

Furthermore, the data used for this project was also retrieved from the same source of related work 3 (ZNSM-iDB[11]), which includes digital images of Ziehl-Neelsen stained Tuberculosis Bacillis taken under microscope.

At the end, the evaluation methods used by this research are accuracy, sensitivity, and specificity. After applying the approach taken by this research, it produced average values of 90.40% accuracy, 93.46% sensitivity, 96.36% specificity.

Chapter 3:

Material and Methods

3.1 Materials

- Keras: an API utilized as a deep learning framework with a focus on making it simple and quick to execute new experiments. Everything from data administration to hyperparameter training to deployment management is covered by the machine learning workflow. It is used on top of TensorFlow.
- OpenCV: (Open-Source Computer Vision Library) is an open-source computer vision and machine learning software library which was created to give computer vision applications a common infrastructure. This library contains more than 2500 highly optimized algorithms that may be used to track camera movements, track moving objects, human behaviors classification , object identification, and detect and recognize faces.
- Python3: is an open-source scripting language that is also high-level, interpreted, interactive, and object-oriented. It was created to be very readable. It has fewer syntactical structures than other languages and typically employs English keywords rather than punctuation.
- Scikit-Learn: is the most useful and dependable machine learning library for Python. It provides a number of efficient tools for statistical modelling and machine learning, such as classification, regression, clustering, and dimensionality reduction, through an uniform user interface. This library was written primarily in Python and is based on NumPy, SciPy, and Matplotlib.
- TensorFlow: The high-level Keras API is used to construct and train models, making it simple to get started with TensorFlow and machine learning. Models can also be created and trained without losing performance or speed. With tools like the Model Subclassing API and the Keras Functional API, it gives users the flexibility and control they need to build complicated topologies. Additionally, it enables quick debugging and simple prototyping.

3.1.1 Data

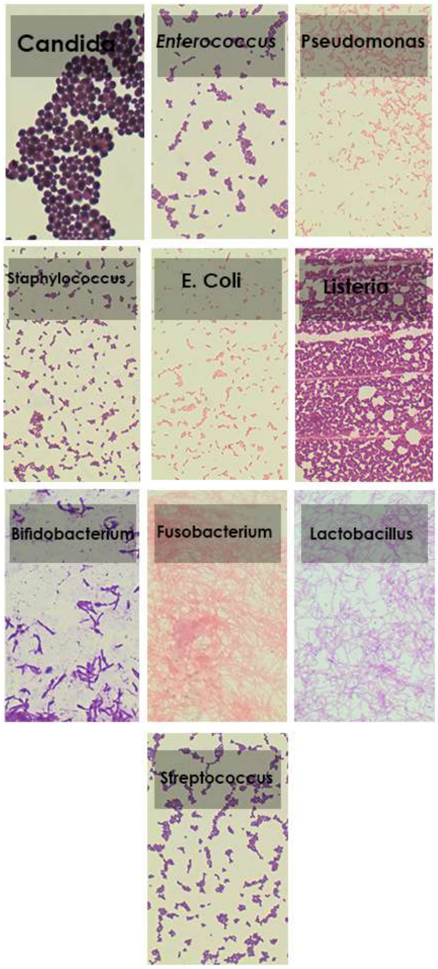

Figure 5: Dataset Sample

*Source (*http://misztal.edu.pl/software/databases/dibas/ *)*

This project dataset is retrieved from DIBaS dataset which contains the images of 10 different types of bacteria and their subspecies taken under the microscope (Figure 5). These images were taken after applying a gram stain test for easy identification. Each type of bacteria is approximately 125mb of size and contains .TIF format images. A custom python code was used to convert the images into .JPG format images for easier use. http://misztal.edu.pl/software/databases/dibas/

3.1.2 Tools

- Anaconda: a Python and R distribution with the aim of simplifying deployment and package management for data science, machine learning applications, big data processing, predictive analytics, etc.
- Google Collaboratory Code: an open-source notebook environment that enables programmers to easily work on and integrate their cloud-based applications.
- Juniper Notebooks: an open-source web tool that allows users to create and share interactive documents featuring computations, graphs, code, etc.
- Visual Studio: To produce GUI (Graphical User Interface), console, Web, Web, mobile, cloud, and other applications, Microsoft developed Visual Studio, an Integrated Development Environment (IDE). With the help of this IDE, managed and native code can be produced.

3.1.3 Environment

- Anaconda Environment: It offers the ability to make environments with multiple versions of Python and/or packages installed and then export, list, remove, and update those environments. It is helpful since it is as if it provides different clean installs in the same device specifically for a certain project.
- Local CPU, intel i7 processor with 6 cores.

3.2 Methods

3.2.1 System architecture Overview

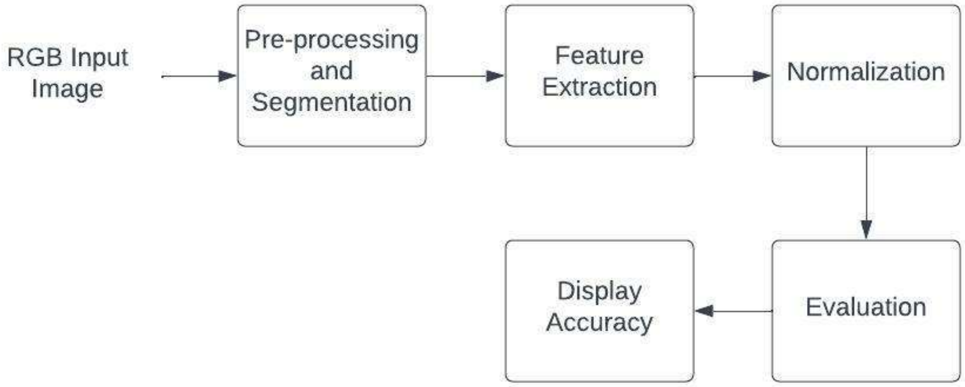

Figure 6: System Architecture Overview

As shown by Figure 6, the approach that was proposed in this project consists of 5 main phases, pre-processing, feature extraction, normalization, evaluation, and display accuracy. The pre-processing phase consists of loading RGB digital images from the dataset acquired, and apply different image processing techniques i.e., edge detection, color thresholding, image segmentation, etc. Furthermore, feature extraction techniques like HOG or SIFT will be applied during the second phase, which will then be normalized before evaluating. In addition, the evaluation process consists of using the normalized data as input for a CNN model, which uses accuracy as an evaluating factor. Lastly, the accuracy will be displayed to determine if the results are acceptable or not. The goal is to use machine learning with the help of different image processing and feature extraction techniques to reach a better accuracy for microorganisms identification.

Chapter 4:

System Implementation

4.1 System Development

This project was developed on different phases in order to assure the final product to possess both good effectiveness and efficiency. Furthermore, different platform and techniques were used in order to guarantee effectiveness and efficiency.

With the help of scikit-learn, the first model was created by using only 7 classes out 10 to check the capability of the machine to correctly classify those groups. Moreover, due to the dataset retrieved consisted of pictures with fixed size, resizing was not needed. However, because of the humongous size of the images, it was still decided to resize the images to 1/3 of their original size. Figure 7

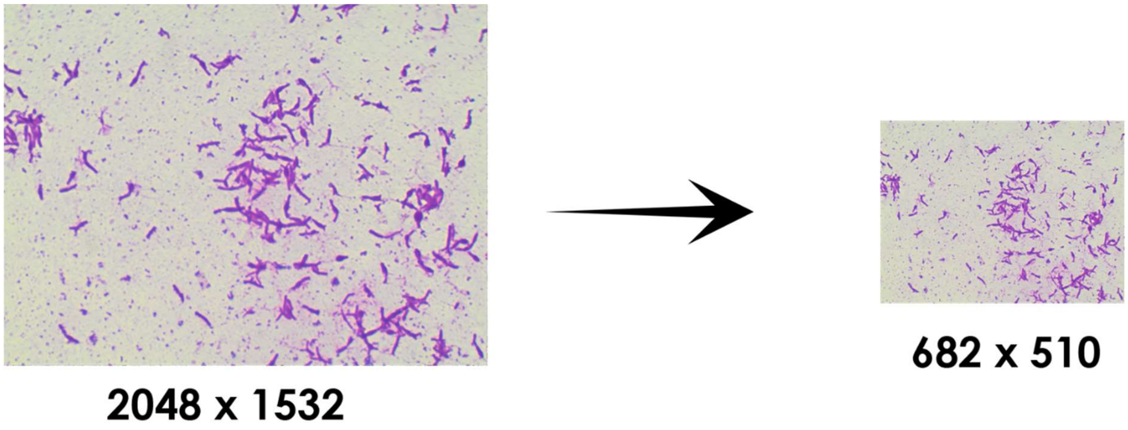

Figure 7: Input after resizing

Hence why, OpenCV library was used because of its good reputation with managing images. As was mentioned above, this project was divided into different phases to assure efficiency, that is why another solution to increase speed was proposed. Since the dataset used in this project consisted in colored images with 3 channels, that will put a load on the processing phase, consequently it was decided to work on a single channel, in other words the grayscale technique was used on the dataset. (Figure 8)

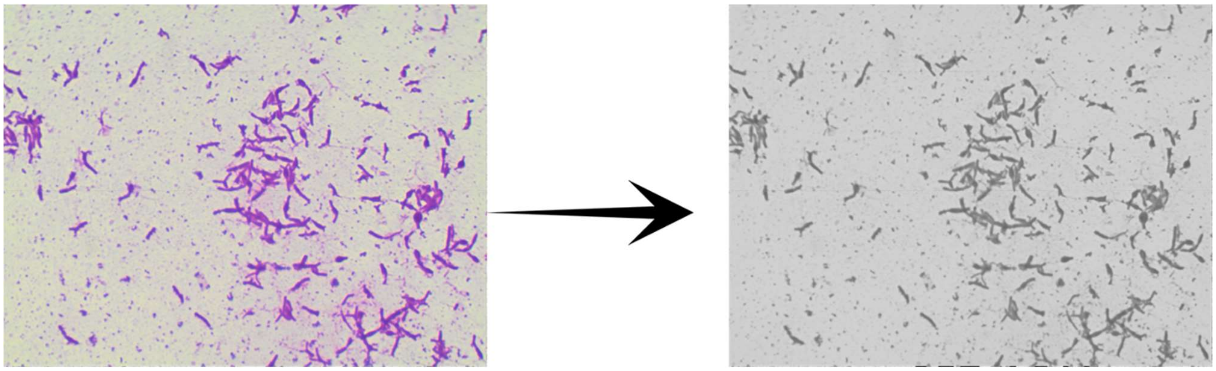

Figure 8: Input before and after applying grayscale technique

Another technique used was masking, specifically background removal. Essentially, the masking techniques used consists of extracting and image from the original according to HSV color ranges.

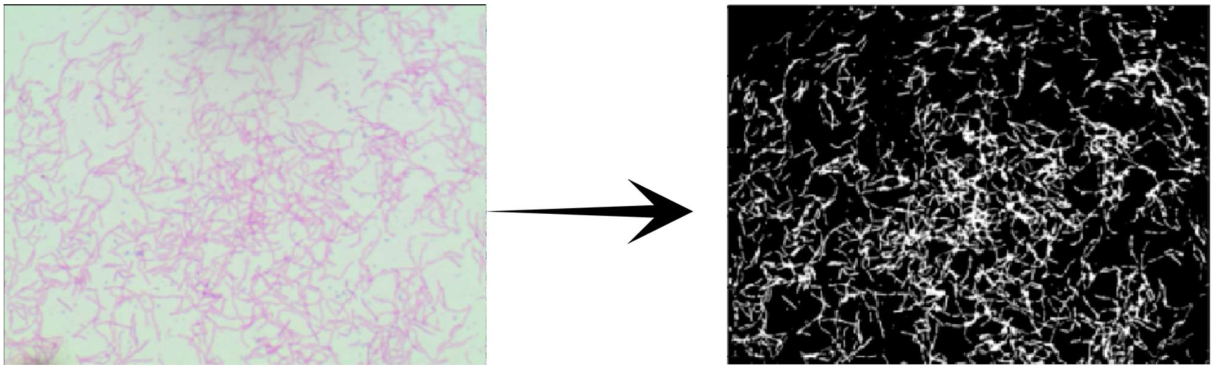

Figure 9: Input before and after applying masking technique

Hue Saturation Value (HSV), a method for mathematically describing color characteristics. A region's primary color is determined by its hue, and its level of color saturation is determined by its brightness. (T. Rehenuma, G. Alim et al., 2010). HSV was applied as a color thresholding approach, as seen in (Figure 9), to eliminate the redundant information in the image and focus on the bacteria itself. In addition, two more models were created next, however this time TensorFlow was used. From the models created, one was a CNN model and the other an ANN. Another technique was used to further improve the results, which consisted of extracting features from the images. Three different approaches were taken to tackle this problem, which are: HOG, SIFT, CNN output after feature extraction. After several tests were taken the third approach developed the best result, hence why another model was proposed.

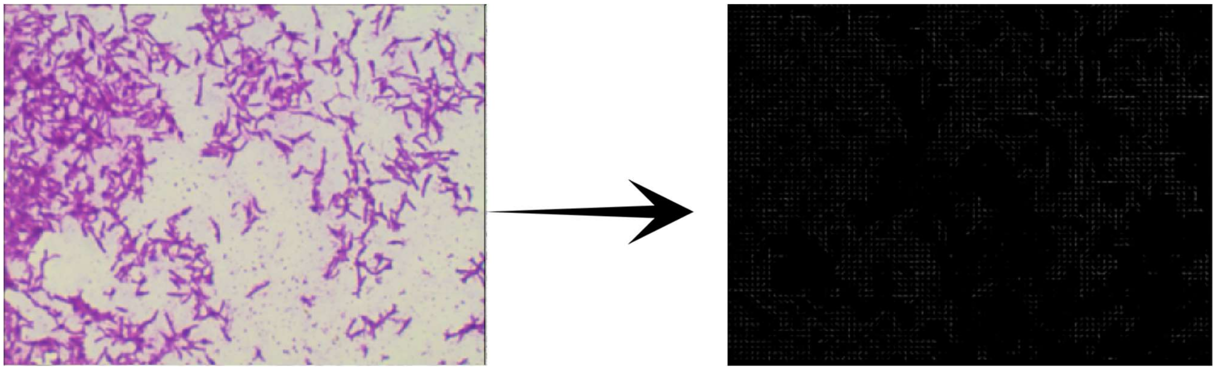

Figure 10: Input before and after applying HOG features extraction

The last model proposed was a CNN-ANN hybrid created by the help of TensorFlow. While being a creative approach to classify and identify bacteria, it also produced great result. The CNN-ANN hybrid was able to achieve the predetermined restrictions made while planning the implementation of this project and was able to achieve the nest results among all models created by great margin.

4.2 System Structure

4.2.1 System Overview

This system consists of 4 main stages, which are: Pre-processing, Feature Extraction, Evaluation, and Displaying Accuracy. In addition, as shown by Figure 11, the first stage is further divided into three different steps. The first step is loading images and their corresponding labels. Essentially, an object will be created consisting of np array of images, and their representing int array labels. Furthermore, the newly created object is divided into four parts, which are: X_train, Y_train, X_test, Y_test. In the division mentioned above, X represents the images and Y represents the labels. Moreover, the test set and train set are randomly chosen with the restriction of 20% being the size of the test set. The next step consists of image resizing, which as mentioned previously it was decided that the images will be rescaled to 1/3 of their original size. Lastly, the last step of pre-processing is data augmentation, which consists of using image processing techniques to aid the system either with speed or effectiveness. The image processing techniques used are grayscale and HSV-masking.

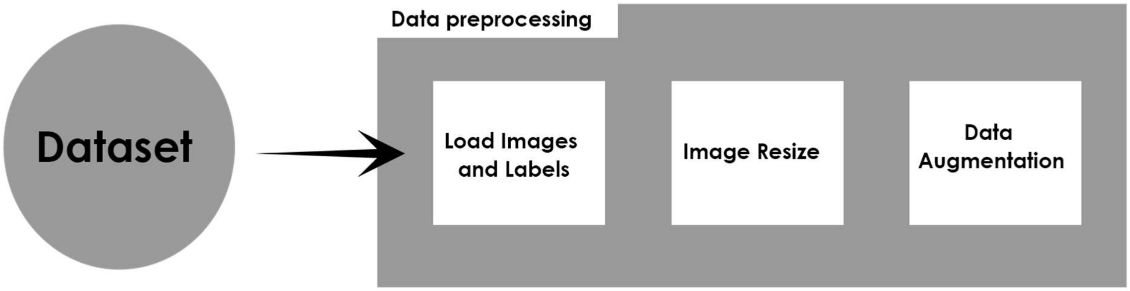

Figure 11: System preprocessing approach

Furthermore, as mentioned previously in this research, feature extraction is the second stage. It can be seen from Figure 12, that three different techniques were used throughout this project, and HOG feature extraction is one of them. HOG, Histogram of Oriented Gradients, is a feature descriptor based on analyzing correctly normalized local histograms of grid-dense image gradient orientations. Even without perfect knowledge of the matching gradient or edge positions, local object appearance and shape can frequently be defined rather well by the distribution of local intensity gradients or edge orientations (Figure 10). (M. Panca, Rahmadwati et al., 2015). Hence why, due to its utility HOG was used in this project to extract features from the images of the dataset. In addition, SIFT is another technique used in this project. SIFT essentially extracts keypoints of objects from original images, and then compares every individual feature with a new image to find matching features based on Euclidian distance. This project implemented SIFT by using the keypoints produced to extract the features from within so that they can be used for the third stage. As stated by M. Yuvaraju, S. K. et al.(2015), Scale Invariant Feature Transform, or SIFT, extracts features that are invariant to scaling, rotation, and image transitions, as well as to some degree to changes in illumination and 3-D camera viewpoint. Object detection and recognition are two of the several applications where it is commonly employed. Consequently, this statement reinforces the idea of using SIFT for bacteria recognition and identification. The last technique used which is also the most successful one, is CNN-feature extraction. Rather than the mundane CNN classification, in this project CNN was used for feature extraction. As CNN already includes segmentation and feature extraction in its process, it was proposed as an approach for feature extraction. Essentially, after the CNN finishes the feature extraction step it is stopped, and its output is extracted into a NumPy array.

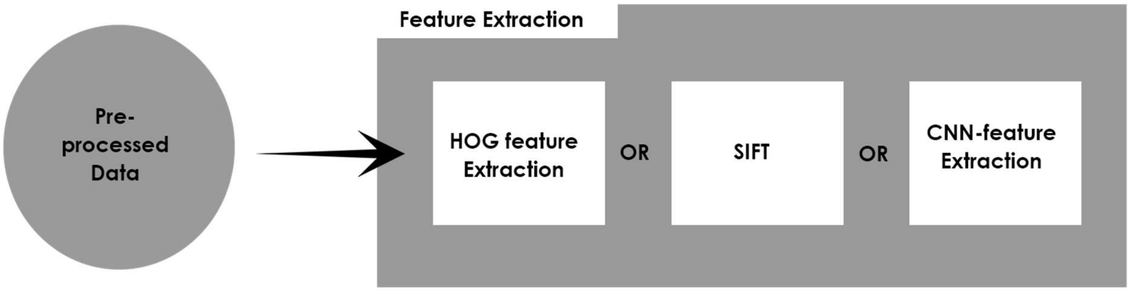

Figure 12: System feature extraction approach

Moreover, the third stage of this project is evaluation, which essentially is the creation of models. Mainly, four different models were created in this project, which are: SKlearn-SVM, CNN-SGD, ANN-Adam, CNN-ANN hybrid. As the last model was created, CNN-ANN hybrid, a peculiar discovery was made. Data augmentation was not essential anymore since the grayscale and other techniques were not used. In addition, CNN-feature extraction was way faster compared to HOG which also contributed to speed acquired by skipping data augmentation. Consequently, the CNN-ANN hybrid achieved the best model rank regarding speed and results.

4.2.2 TensorBoard

Scalars:

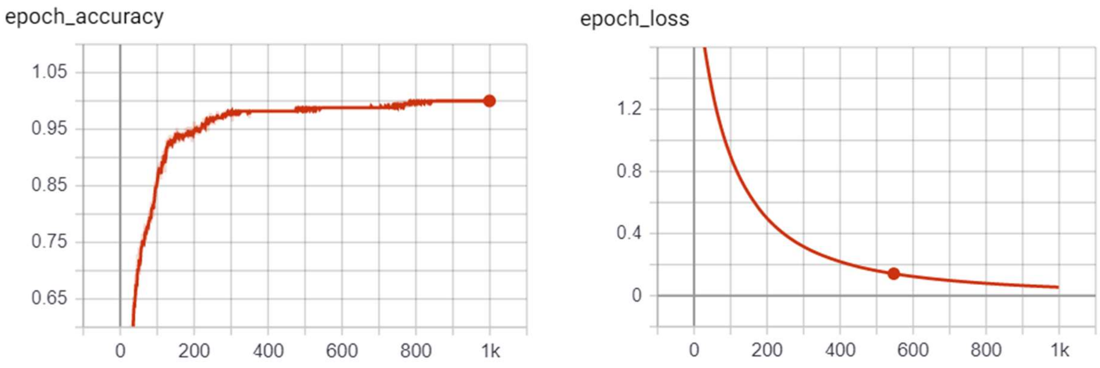

Figure 13: System best-case Scalars

Graphs:

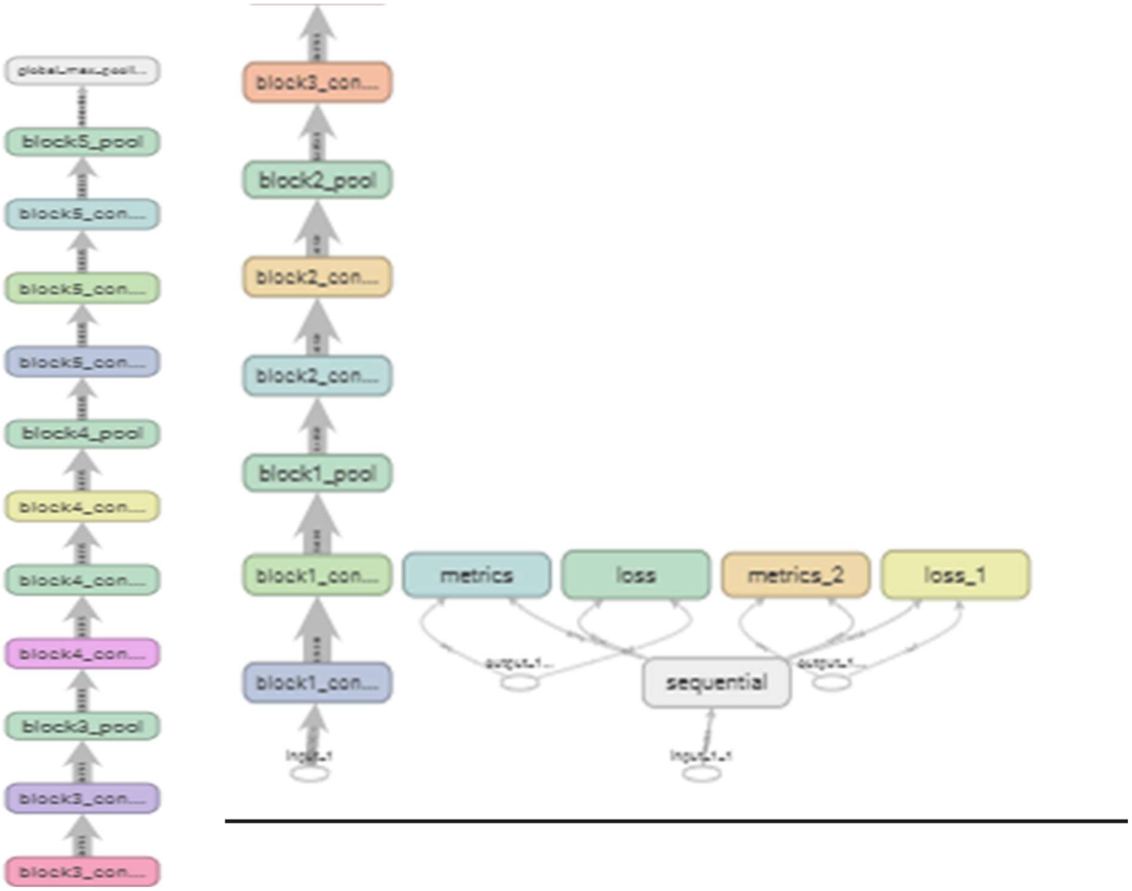

Figure 14: System best-case Graph

Distributions:

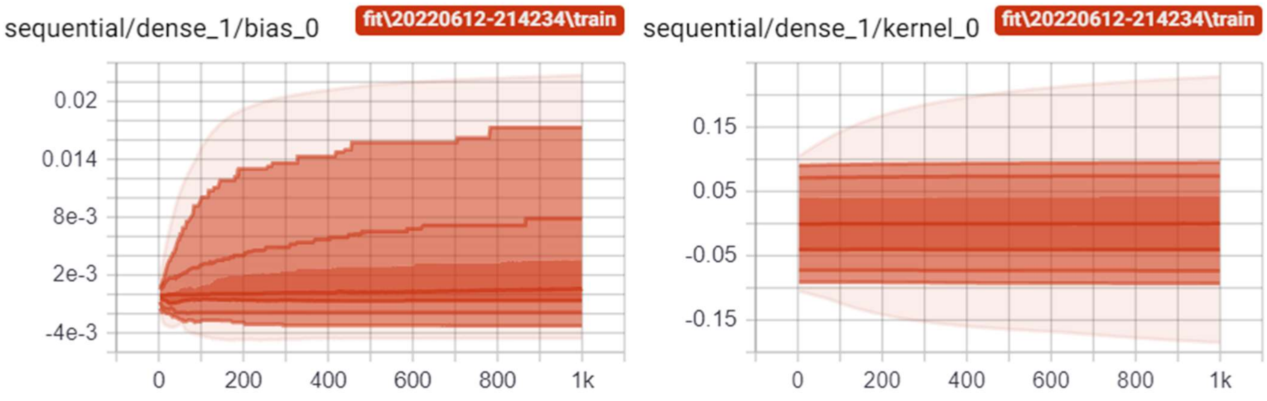

Figure 15: System best-case Distributions#1

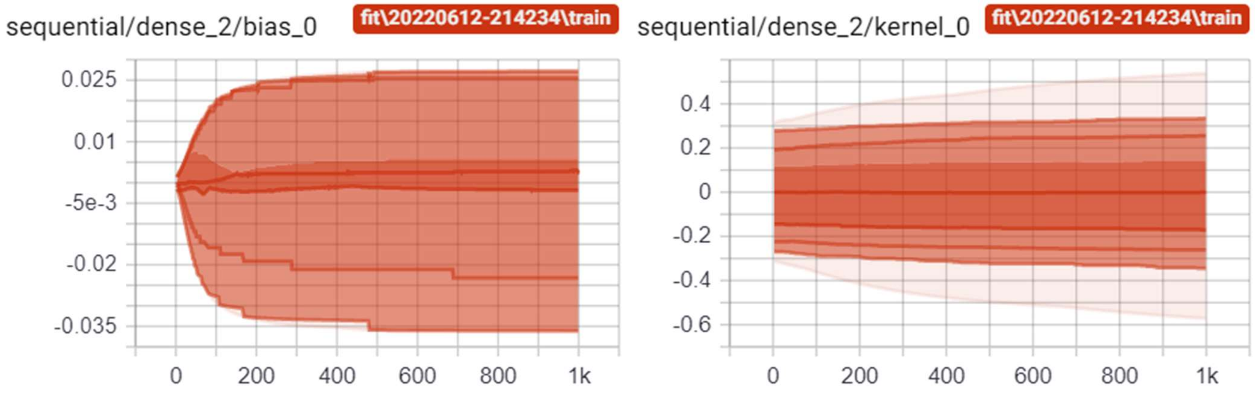

Figure 16: System best-case Distributions#2

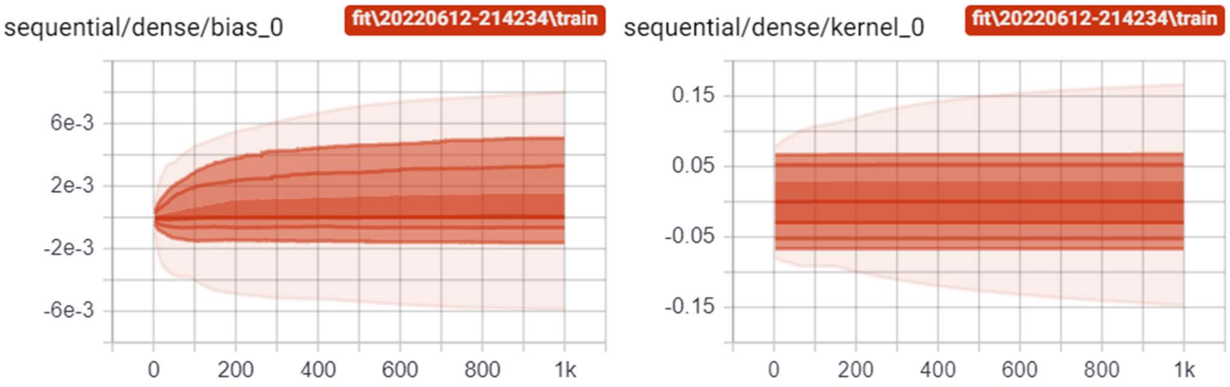

Figure 17: System best-case Distributions#3

Ah shown by the previous distribution it was determined how the weights behaved during training, and the number of weights that were actually activated. Distributions#2 (Figure 16) showed to be the one that has the most weights activated, followed by distributions#1 (Figure 15) and distributions#3 (Figure 17).

Histogram:

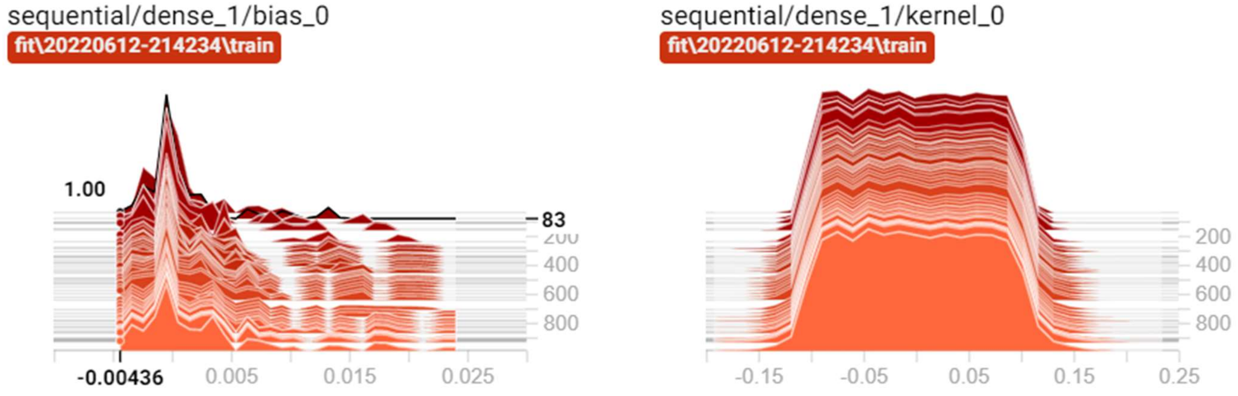

Figure 18:System best-case Histograms#1

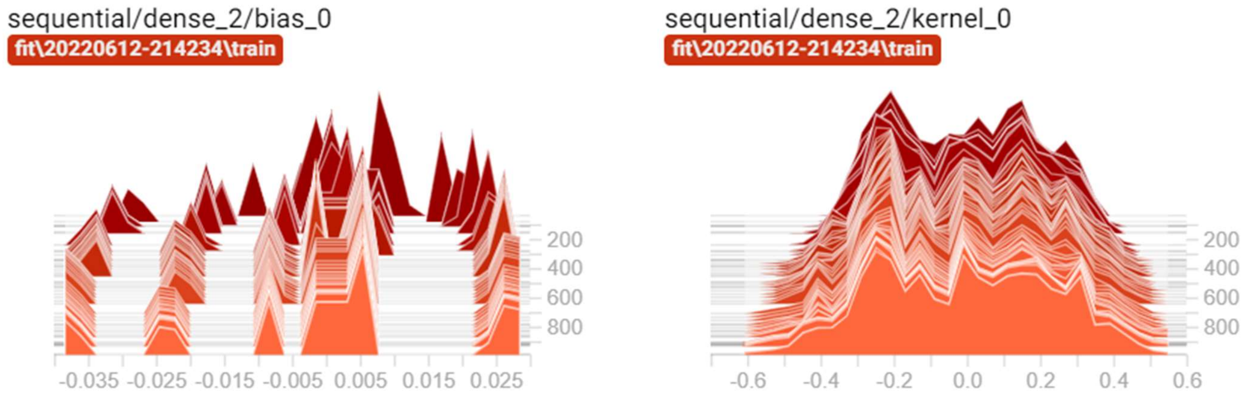

Figure 19:System best-case Histograms#2

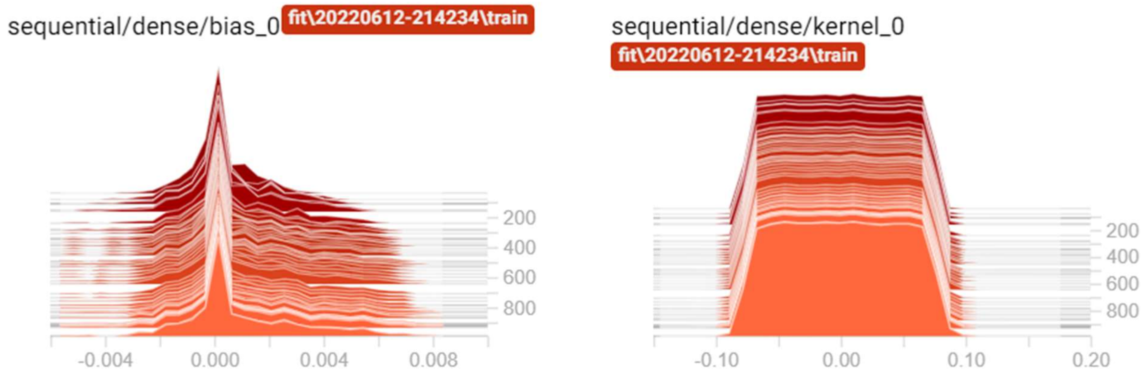

Figure 20:System best-case Histograms#3

Furthermore, the previous histograms show how the values change while training at different times and at different layers. Moreover, it can be determined that the kernel is the most responsible for changing the values throughout the training process. (Figure 18, Figure 19, Figure 20)

4.3 System Running

4.3.1 SciKit-Learn-SVM Model

This model consists of several functions, which are: grayscale, resize, hog. As the system was being developed it was discovered that grayscale and resize could be done simultaneously with the same function. Since the image consists of 3 channels, dividing the shape of the image by a factor of three results in both resizing of the image and converting it to grayscale. (Figure 21)

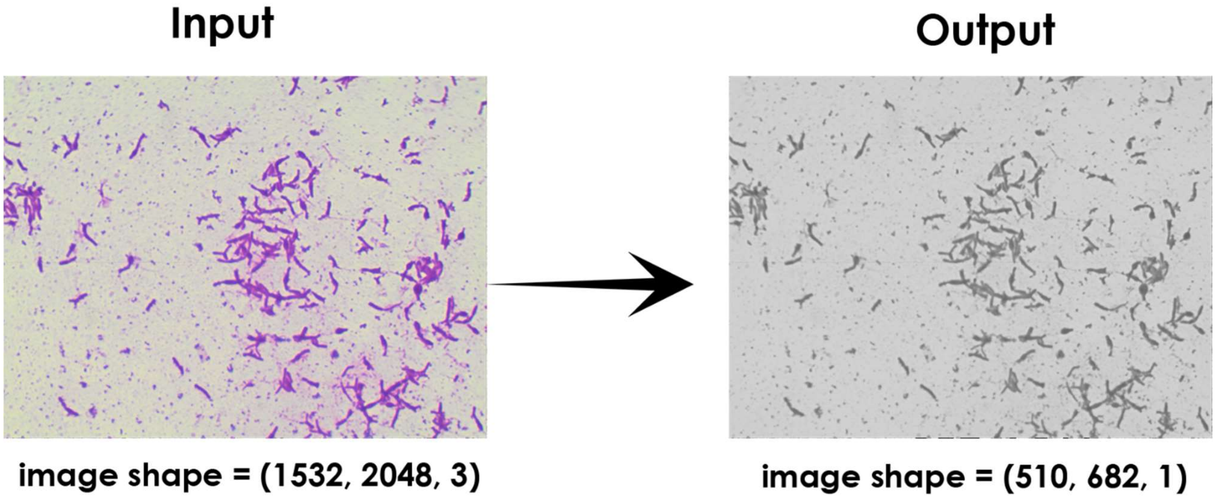

Figure 21: Image shape after rescaling and applying grayscale

The next function used in this component is HOG, also known as Histogram of Oriented Gradients. This function consists of using HOG as a mean to extract features from an image which further decreases the image shape. (Figure 22)

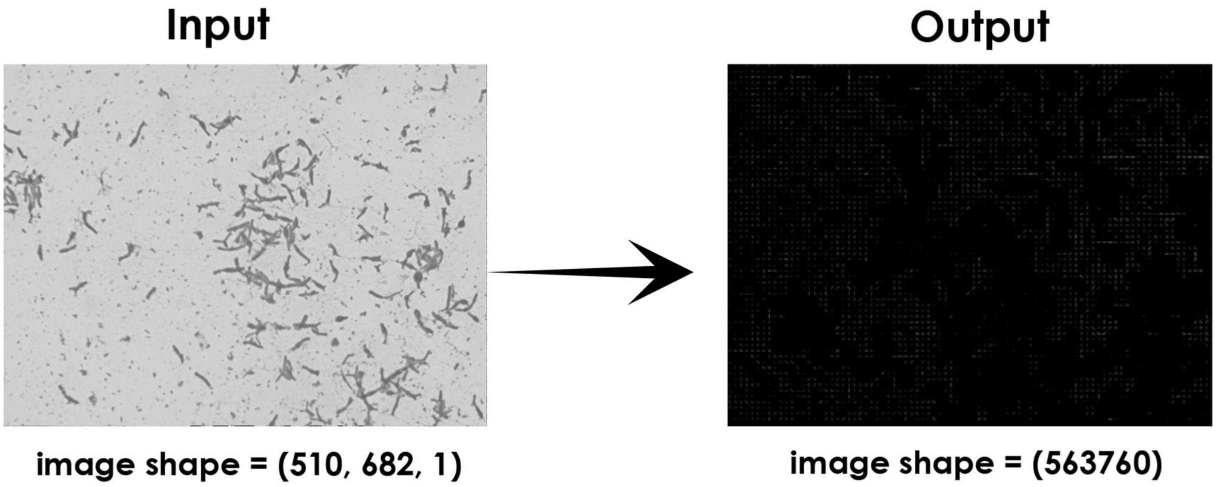

Figure 22: Image shape after applying HOG algorithm

4.3.2 CNN-Basic Model

This model can be called an extension to the Scikit-Learn Model. It was the implementation of the rescaling and grayscale techniques used from Scikit-Learn with the help of TensorFlow. After the pre-processing techniques were done. It was decided to skip the feature extraction techniques due to it being a function already inside the CNN. In other iterations of this model other techniques were applied to this model, but it was decided to move to another approach due to the high computational times produced by this model.

4.3.3 ANN-Adam Model

This model consists of mainly of two functions, background removal (masking) and SIFT. As previously mentioned in this research, background removal is implemented by extracting HSV color ranges from an image. Since the dataset used only includes 2 color concentrations, pink and purple, and HSV range containing these 2 colors was used. (Figure 23)

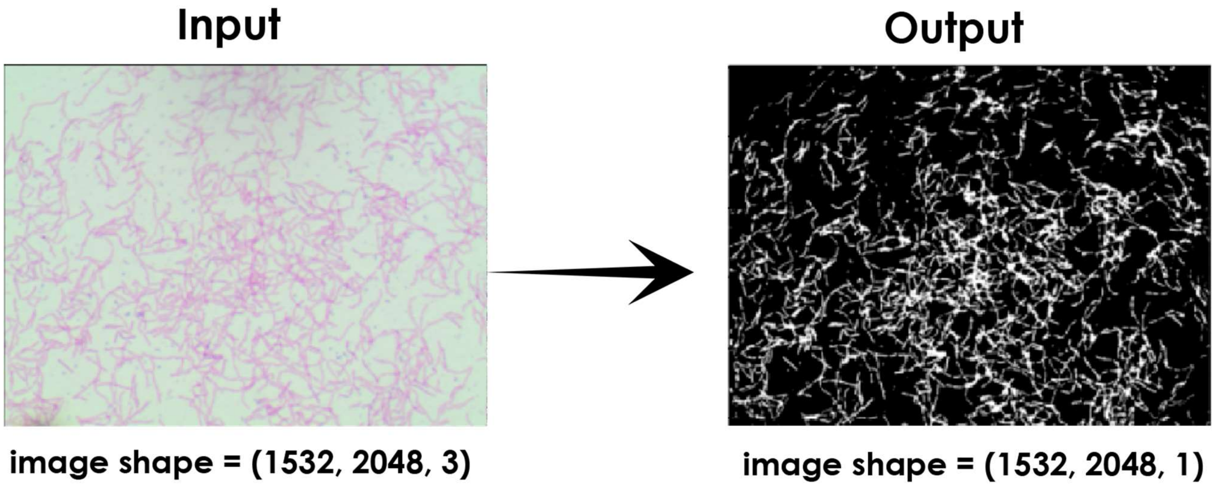

Figure 23: Image shape after HSV color-thresholding

Another function used by this model is SIFT. Essentially, SIFT extracts feature descriptors from images that can later be compared with other feature descriptor to calculate the similarities with Euclidean distance. However, in this model SIFT was used to produce features from the feature descriptors derived from SIFT. Consequently, applying this function changed the shape of the input. (Figure 24)

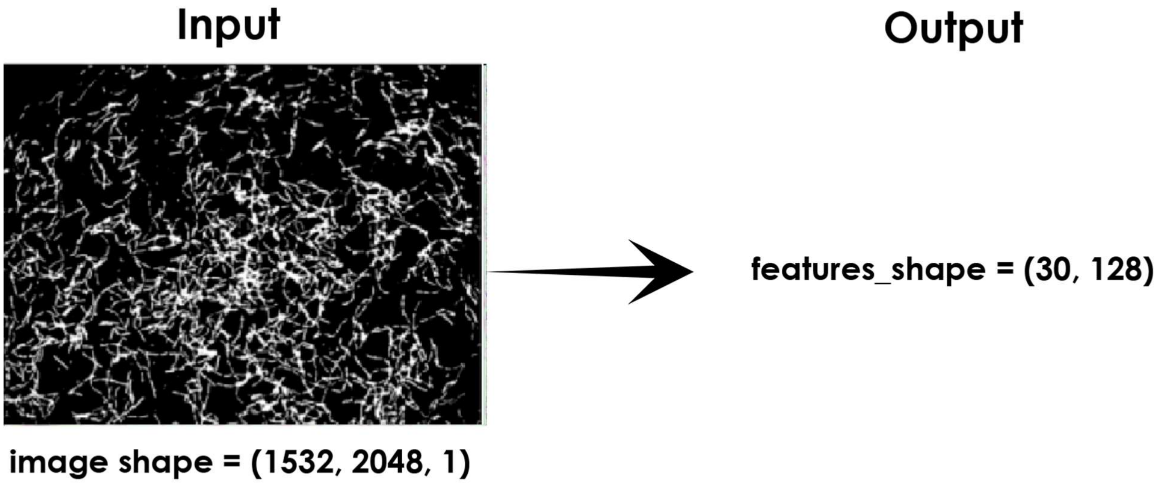

Figure 24: Image shape after applying SIFT algorithm

4.3.4 CNN-ANN Hybrid Model

This model consists of two functions, Resize and CNN-feature extraction. The resize function used in this model is different from the one used in the SciKit-Learn model. Essentially, rather that reshaping the image shape to achieve both resizing and gray scaling, this resize function resizes the image by a factor of 6.8. (Figure 25)

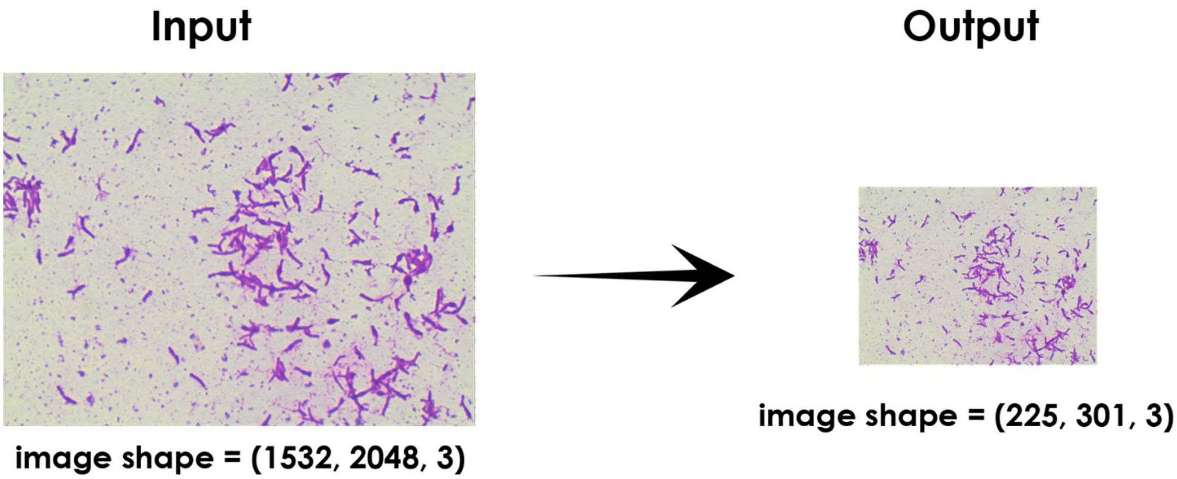

Figure 25: image output after resizing by a factor of 6.8

Lastly, the second function used by this model is CNN-feature extraction.
This function makes use of a creative idea, which consists of using a CNN model to segment and extract features from images rather than using other techniques. Consequently, this makes the job easier since is way faster than the other options. (Figure 26)

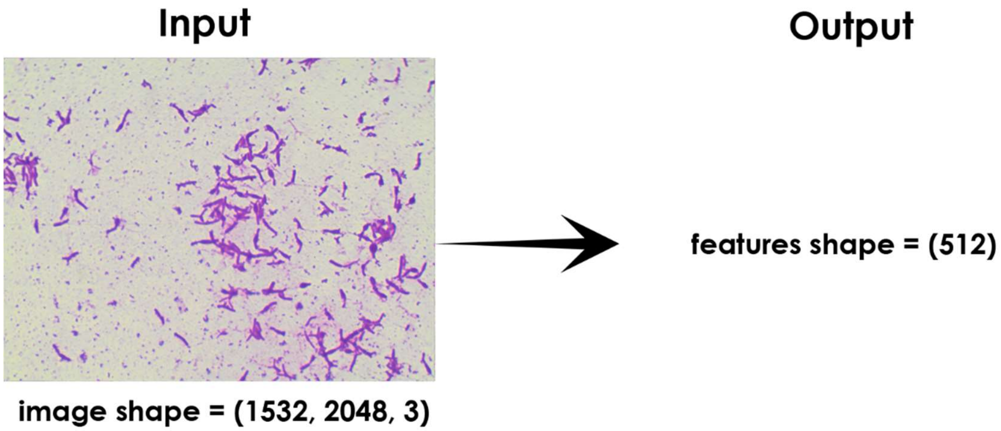

Figure 26: Input after applying CNN for feature extraction

Chapter 5:

Results and Evaluation

5.1 Testing Methodology

First of all, from the multiple methods that can be used to evaluate the performance of the model used to correctly classify and identify bacteria, accuracy and precision are the methods used in this project. Furthermore, the dataset is split into training data and testing data to further maximize the effectiveness of the model. As a result, the testing data will consist of 20% of the original dataset size and the remaining 80% will be for training.

5.2 Results

5.2.1 Best Results Case

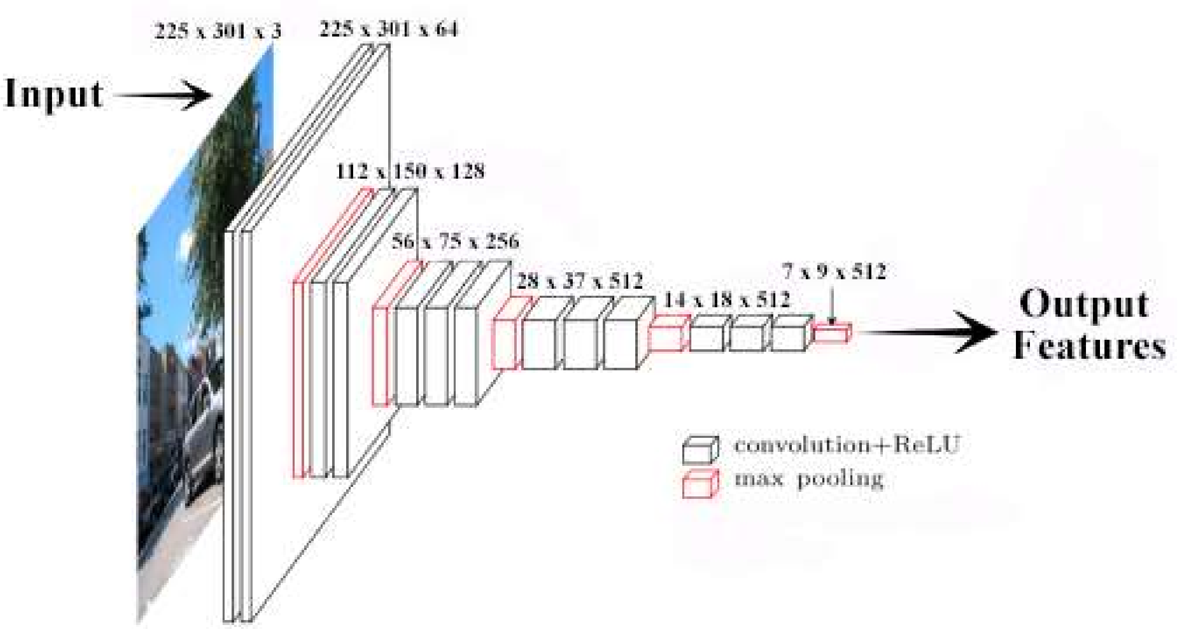

Figure 27: CNN Feature Extraction Architecture

The best model was a hybrid between CNN and ANN, and the structure was divided into CNN to perform feature extraction and the resulting output was inputted to an ANN. As shown by Figure 27, the CNN feature extraction model was generated by using 5 convolutional layers with their corresponding ReLU activation function and 3x3 kernel. Previously, techniques as SIFT for feature extractions and other image segmentation techniques were suggested to enhance the results of this research, however they were recognized to be redundant since CNN is capable of such features. Consequently, it was opted to use the CNN for that purpose and use it in conjunction with ANN to further maximize the results. In addition, the classifier used in ANN was SGD (Stochastic Gradient Descent). Moreover, the model was trained for 1000 epochs and batch size of 1, which successfully identified almost every bacterium correctly from the validation set.

5.2.2 Average Results Case

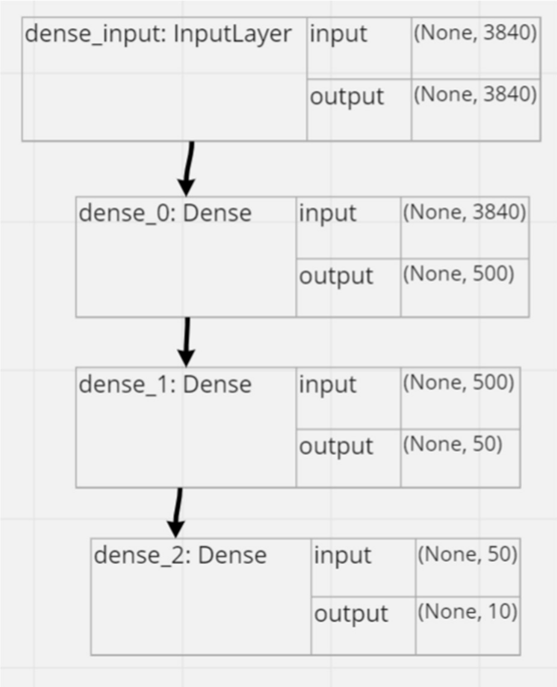

Figure 28: ANN (Average Model) Architecture

The average model was generated by using an ANN architecture which is shown in Figure 28. As shown by the previous figure, the ANN architecture consists of 3 dense layers, first two layers use ReLU activation function, while the third layer uses Softmax. Many combinations of techniques were used in conjunction to this model, However, using SIFT alone produced better results while maintaining the high computational speeds offered by ANN. Consequently, the model was trained for 100 epochs and batch size of 1 with Adam optimizer. As a result, this model showed to be a great improvement to the worst-case model in both speed and accuracy.

5.2.3 Worst Results Case

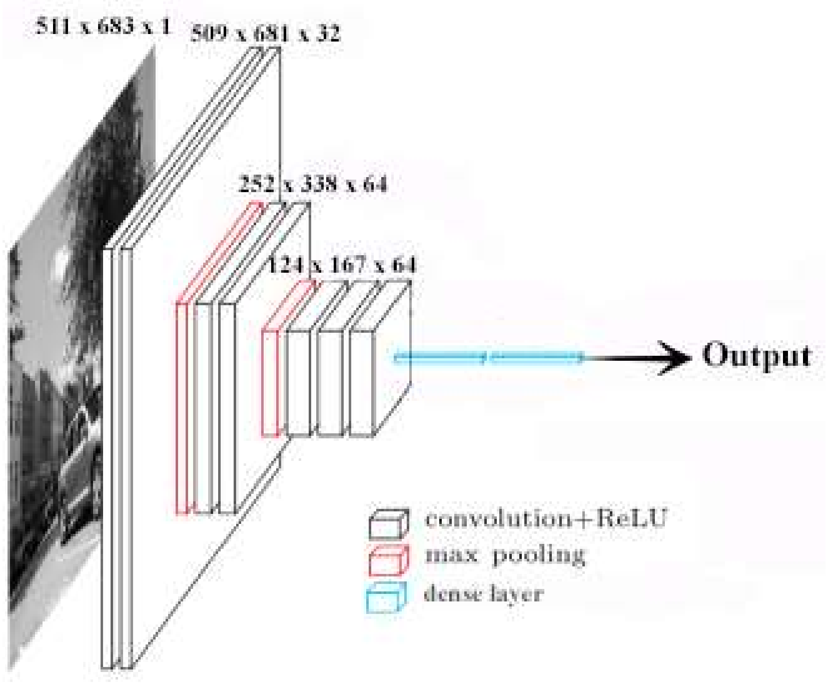

Figure 29: CNN (Worst Model) Architecture

The worst model was a basic CNN without much preprocessing done to the images other than resizing and gray scaling. As shown by Figure 29, the current CNN model was generated by using 3 convolutional layers with their corresponding ReLU activation function and 3x3 kernel. This model produced terrible result and was not even able to endure a batch size of 2 without crashing. Hence why, the model was trained for only 10 epochs and a batch size of 1, which demonstrated to scarcely differentiate the bacteria from the validation test.

5.2.4 Limitations

The selected dataset consisted of different types of bacteria to be further identified by a model developed during this research. As much as the results produced a positive outcome, the number of images may become inadequate in a larger scale scenario. A solution for this issue was using data augmentation, but it would result into a loss of quality to the images decreasing the accuracy. These limitations were produced because of the difficulty to find free datasets for bacteria due to them being from the scientific department. If investments are made, this project could achieve greater heights.

5.3 Evaluation

5.3.1 Accuracy Evaluation

Table 1: Summary of the Classification Models

Models Number	No. of Conv layers	Kernel size	Activation Function	Batch Size	Epochs	Avg. Time / sample	Validation Precision	Validation Accuracy
Scikit-Learn SVM	None	None	None	1	1000	Not determined	30.95%	71.90%
Basic CNN	3	3□3	ReLU	1	10	525ms	23.07%	36.84%
CNN-HSV	2	2□2	ReLU	1	100	20ms	13.79%	55.71%
ANN-SIFT	None	None	ReLU and Softmax	1	100	5ms	57.14%	83.87%
CNN-ANN Hybrid 1	5	3□3	ReLU	1	100	3ms	85.71%	94.53%
CNN-ANN Hybrid 2	5	3□3	ReLU	1	1000	140us	96.8%	98.12%

Accordingly to what is shown in Table 1], different implementations for the identification of bacteria where made. Moreover, different model architectures were used to ensure the best output to be accomplished. Furthermore, changes in the parameters like the convolutional layers and epochs were made in contribution to the different preprocessing techniques applied throughout the project. As a result, the CNN-ANN hybrid model resulted with the best accuracy and precision among all the models used by a great margin. After running the model with 100 epochs, it resulted with a precision of 85% and accuracy of 84%. Furthermore, this model was also run with 1000 epochs, which resulted in best accuracy so far with a precision of 95% and an accuracy of 98%.

As shown by Figure 30, the accuracy and loss of both training and validation according to the results given by the best-case model (CNN-ANN Hybrid). As shown from the first graph, it can be determined that the best model produced 99% training accuracy and 95% validation accuracy. Furthermore, the second graph represents the training and validation loss produced by the best-case model, which resulted in 0.05 training loss and 0.18 validation loss.

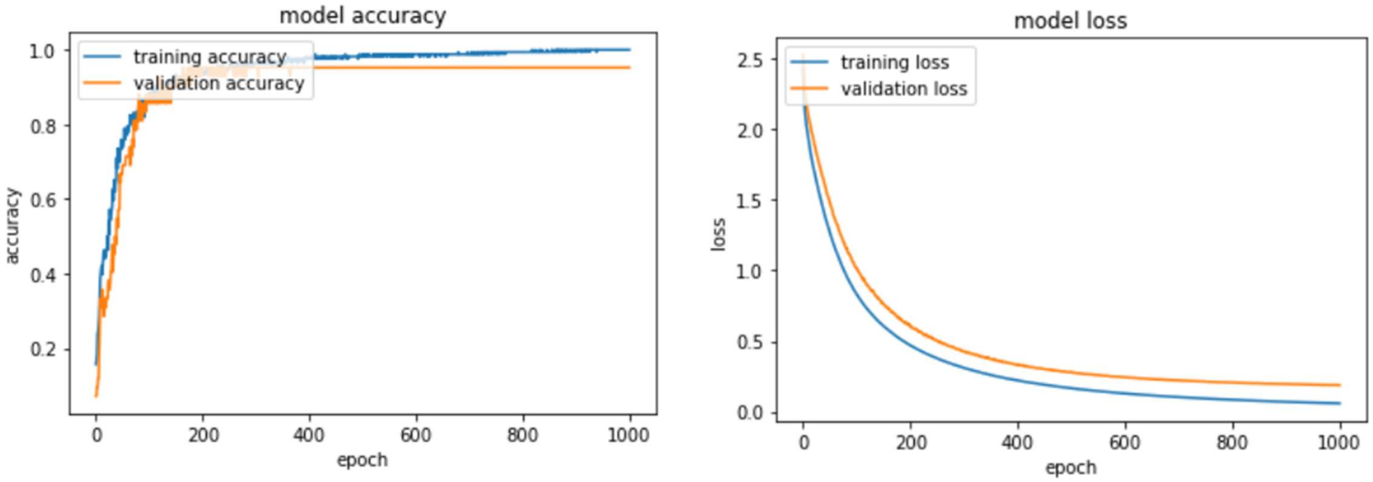

Figure 30: Graphs of the best CNN model loss and Accuracy

5.3.2 Time Performance

Regarding the time, using the local PC CPU was considered a disadvantage, in which this idea is strengthened after the worst-case model was developed. This model took 525ms per sample and 27s per epoch. It was recommended to use Google Colab GPU to tackle this problem, but the aim of this project was to provide a lightweight service. As a result, some performance improvements were suggested during this project, and the CNN-ANN hybrid successfully solved this issue. This new model provided surprising speeds, 2ms per sample and less than 1s per epoch. Consequently, the significant advantage of Google Colab became inessential.

Chapter 6:

Conclusion and Future Work

6.1 Conclusion

In this research, a fully automated system was developed by using different algorithms capable to identify and classify different bacteria among many types. After several approaches were tested, it was determined that the best option was to make use of deep learning and its different techniques to solve the problem this research is responsible to solve. All the decisions made during this project had the sole purpose of building a successful system able to recognize different bacteria while maintaining high accuracy standards. Consequently, it was decided that in order to tackle the problem proposed by this research, that an algorithm able to distinguish and correctly classify different objects by using images was required. Hence why, at the beginning the Convolutional Neural Network algorithm was selected. However, due to some difficulties it was decided to take another approach. Furthermore, a library called Sci-kit Learn was used to aid in recognizing the best structure to solve this problem. Sci-kit Learn was useful for understanding and coming up with ideas thanks to its simple and fast model creation and training. Moreover, after trying different techniques with Sci-kit Learn, another approach was tested. Artificial Neural Network, also known as ANN, was confirmed to produce faster results while still competing with CNN's precision by using different pre-processing techniques. Finally, after researching more about the different algorithms used during this project, it was suggested to make use of the best aspects from both ANN and CNN. Hence why, a hybrid model was proposed which makes use of CNN's precision with images and ANN's speed. Consequently, such approach proved to exceed all previous approaches, thus it was selected as the final model for this project.

6.2 Problem Issues

6.2.1 Technical issues:

At the beginning of development, the extreme size of the images from the dataset caused many crashes due to RAM usage and CPU usage. Such issue was first tackled by turning the dataset into greyscale images since it greatly decreased the input shape of the images. However, this produced further issues due to the importance of colors in the identification of bacteria, thus producing terrible results. As a result, the approach that proved to be the most useful was rescaling the images. However, this produced long computational time as other pre-processing techniques were used at the same time. Lastly, it was determined that some redundant algorithms were used in the pre-processing part which were all solved after the idea of the hybrid model was introduced.

6.2.2 Scientific issues:

While choosing a batch size for training the model, any value greater than 1 caused severe crashes. This was caused for the same issue mentioned previously in this project, which was the huge size of the images (1532x2048). After some changes were made a batch size of 8 was possible without crashing but resulted in slower training speeds. As a result, it was decided to keep the batch size with a size of 1 to keep from producing unnecessary load to both CPU and RAM. This issue could have been solved by using GPU rather than CPU, but the goal of this project was to create a lightweight model capable to work while still maintaining high accuracy standards.

6.3 Future Work

In future works, is recommended to work on a bigger dataset to further increase both the research usefulness for identifying more types of bacteria and increase the variety of the images for the different types of bacteria to further increase precision against abnormal data. Furthermore, even if this project was developed by making use of only CPU for the lightweight goal in mind, further tests with GPU or cloud services like AWS could prove to be useful for a larger scale project. In addition, tweaking the structure of the models used may potentially produce better results than those that were generating during this research.

References

[1] Anderson, James & Dixon, Dennis & Maloy, Stanley & Nacy, Carol & Tiedje, James. (2007). Basic Research on Bacteria: The Essential Frontier.

[2] B. Zieliński, A. Plichta, K. Misztal, P. Spurek, M. Brzychczy-Włoch, and D. Ochońska. Deep learning approach to bacterial colony classification. *PLOS ONE*, 12(9), 1-14, 2017.

[3] Chiang, Chia-Yen & Barnes, Chloe & Angelov, Plamen & Jiang, Richard. (2020). Deep Learning-Based Automated Forest Health Diagnosis From Aerial Images. IEEE Access. 8. 1-1. 10.1109/ACCESS.2020.3012417.

[4] El-Melegy, Moumen & Mohamed, Doaa & Elmelegy, Tarek. (2019). Automatic Detection of Tuberculosis Bacilli from Microscopic Sputum Smear Images Using Faster R-CNN, Transfer Learning and Augmentation. 10.1007/978-3-030-31332-6_24

[5] F. Pedregosa & G. Varoquaux & A. Gramfort et al. *Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.*

[6] Grossi, Enzo & Buscema, Massimo. (2008). Introduction to artificial neural networks. European journal of gastroenterology & hepatology. 19. 1046-54. 10.1097/MEG.0b013e3282f198a0.

[7] Gulpreet Kaur Chadha, Aakarsh Srivastava, Abhilasha Singh, Ritu Gupta, Deepanshi Singla. (2020). An Automated Method for Counting Red Blood Cells using Image Processing: Procedia Computer Science. 10.1016/j.procs.2020.03.408. https://doi.org/10.1371/journal.pone.0184554

[8] M, Yuvaraju & Sheela, K & Sobana Rani, Sheela. (2015). Feature Extraction of Real-Time Image Using Sift Algorithm. International Journal of Research in Electrical & Electronics Engineering. 3. 1-7.

[9] Mithra, K.S., Sam Emmanuel, W.R. Automated identification of mycobacterium bacillus from sputum images for tuberculosis diagnosis. *SIViP* **13,** 1585–1592 (2019). https://doi.org/10.1007/s11760-019-01509-1

[10] Mudjirahardjo, Panca & Rahmadwati, & Sulistiyanto, Nanang & Setyawan, Raden. (2015). Multi Orientation Performance of Feature Extraction for Human Head Recognition. Journal of Engineering and Applied Sciences. 10. 10541-10547.

[11] Shah, M.I., et al.: Ziehl-Neelsen sputum smear microscopy image database: a resource to facilitate automated bacilli detection for tuberculosis diagnosis. J. Med. Imaging (Bellingham, Wash.) 4(2) (2017)

[12] Tabassum, Mirza Rehenuma & Gias, Alim & Kamal, Md & Muctadir, Hossain & Ibrahim, Muhammad & Shakir, Asif & Imran, Asif & Islamm, Saiful & Rabbani, Md. Golam & Shah, Khaled & Islam, Md & Begum, Zerina. (2010). Comparative Study of Statistical Skin Detection Algorithms for Sub-Continental Human Images. Information Technology Journal. 9. https://doi.org/10.3923/itj.2010.811.817

[13] Vojislav. (2005). Support Vector Machines – An Introduction. https://doi.org/10.1007/10984697_1

[14] Zhang, J., Li, C., Rahaman, M.M. *et al.* A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches. *Artif Intell Rev* (2021). https://doi.org/10.1007/s10462-021-10082-4