A

MAJOR PROJECT-III REPORT

on

Automated Bacteria Prediction and Morphology Extraction using machine learning model:
Advancing Precision in Infection Treatment

Submitted by:

Dhruv Sahgal
(210283)

under mentorship of

Dr. Devanjali Relan
(Associate Professor)

Department of Computer Science Engineering
School of Engineering and Technology
BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

May 2024

CANDIDATE'S DECLARATION

I hereby certify that the work on the project entitled, "Automated Bacteria Prediction and Morphology Extraction using Machine Learning Model: Advancing Precision in Infection Treatment", in partial fulfillment of requirements for the award of Degree of Bachelor of Technology in School of Engineering and Technology at BML Munjal University, having University Roll No.210283, is an authentic record of my own work carried out during a period from March 2024 to May 2024 under the supervision of Dr Devanjali Relan.

Dhruv Sahgal

SUPERVISOR'S DECLARATION

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name:

Signature:

ACKNOWLEDGEMENT

Table of Contents

ABSTRACT

In order to diagnose diseases and arrange treatments, it is essential to automatically identify and classify bacteria in medical imaging. This research presents a thorough framework for doing so using deep learning techniques. The dataset utilised, consists of medical pictures obtained using a variety of imaging techniques, such as microscopy and medical photos. State-of-the-art object detection algorithms are used to identify bacteria and classify them into several groups according to morphological characteristics particularly with the help of the Detectron2 library. The method involves preprocessing the pictures, adding bounding boxes to them, and training a convolutional neural network (CNN) to detect instances of bacteria. Metrics like mean average precision (mAP), recall, and precision are used in the model's performance evaluation to make sure it is accurate and robust. In an effort provide insight into the model's performance and highlight areas in need of development, the detection results are also graphically shown. The suggested framework demonstrates promising results in terms of efficiency and accuracy, opening the door for further developments in automated bacteria detection systems. By giving medical professionals a reliable instrument to help with the detection and treatment of bacterial infections, this project advances the field of medical image analysis. Bacteria plays an important role in various areas like agriculture and medicine, out of which some bacteria are beneficial to humans but a majority of them pose a significant health risk causing various illnesses across different organisms. However with the large availability of data and technical advances in the field of medical and computer science, the machine learning methods have been widely used and have shown tremendous performance in automatic detection of bacteria. The adaptation of latest technology using different artificial intelligence techniques are greatly helping microbiologists in solving extremely complex problems.

This study carries out a detailed analysis of different machine learning methodologies in the field of bacterial classification along with their limitations and future scope.

LIST OF FIGURES

Chapter 1

Introduction

While certain microorganisms are essential for maintaining soil fertility, producing fuel, fermenting food, treating sewage, making medications, and other processes, many germs are extremely dangerous to people and the environment. Microorganisms, which include bacteria, fungus, viruses, algae, and protozoa, are abundant on Earth and are crucial to both the environment and human existence. Microbes are widely employed in a variety of industries, food processing, agriculture, medicine, and other fields, yet very few of them are good for humans. Since they are too tiny to be seen with the human eye, microorganisms, also known as microbes, are microscopic creatures that can only be viewed under a microscope. There are two types of microorganisms: single-celled and multicellular. Human illnesses such as AIDS, polio, typhoid, food poisoning, cancer, and milder colds are all brought on by bacteria. Microbes are generally highly essential living things on Earth, and humans need to be aware that they pose a serious threat to other living things as well. Microbes Prokaryotic microorganisms, such as bacteria, are characterized by the absence of a proper nucleus and a dispersed cytoplasm that contains their genetic material (DNA). These are categorized as spherical (cocci), spiral (Sprillium), rod (bacilli), comma (vibrious), or crockscrew (spirochaetes) based on their shape. Although they are often everywhere, humans serve as their ideal host. While a large group of bacteria are benign and even beneficial in many ways, some are harmful and cause a variety of illnesses in both people and other animals, including cholera, TB, tetanus, and typhoid fever. Numerous of these may potentially prove to be lethal. Because bacteria are dual in nature, we may divide them into a number of groups based on their critical roles in genetic engineering, biotechnology, food processing, medical diagnostics, and other fields. The classification of bacteria was traditionally done by hand by microbiologists. The most often used methods involve looking at bacterial species' phenotypic traits under a microscope, such as size, shape, color, etc. The entire process of bacteria classification relies solely on human experience, including extensive training for highly skilled operators.

Chapter 2

Problem statement

In this project we are trying to solve the problem of bacteria detection using various machine learning techniques with the help of which microbiologists would not waste their precious time trying to classify the bacteria. Microorganisms are extremely important to human existence and the earth's ecosystem. People may readily observe the significance of bacteria in their daily lives by providing a global example of the COVID-19 scenario. In a similar case, bacteria are regarded as among the most significant microbes and have a significant impact on both the ecology of the planet and also the human existence. Given their tiny size and inability to be seen with the human eye, bacteria are difficult to identify and classify. The microbiologist wastes a lot of valuable time classifying materials using traditional methods since they are extremely time-consuming. Research on bacteria is only conducted if the species in question has been correctly classified and assigned to a suitable category. Because of the little variations in size and form, categorization across families is simpler than that within families. Conventional approaches classify bacteria according to their morphological characteristics, which are less in number than those found in automated technologies that use a contemporary computing environment. Researchers were able to automate the categorization of bacteria based on picture analysis thanks to the introduction of AI in the field of image classification utilizing various machine learning approaches. The microbiologist benefits greatly from the application of various machine learning algorithms, which shortens the time required to classify microorganisms. ML-based classification overcomes the drawbacks of conventional approaches by employing diverse feature extraction and feature selection techniques in place of the finite amount of morphological features used in classic procedures. A decision made using a machine learning model has a very large number of parameters, which increases its accuracy.

Chapter 3

Literature review

Intensive research was done and articles were searched and gathered from famous sources like IEEE Explore, Google Scholar, Research gate. One subset of AI that has excelled in computer science is machine learning. After being trained on a set of data, machine learning algorithms may be used to forecast the results of fresh data. Researchers in microbiology analyze digitized tiny bacterial photos using machine learning techniques. Bacteriology is the study of bacteria; researchers examine the bacteria using their genetic makeup and physical characteristics. In several fields, including food processing, biotechnology, fiber retting, pest management, and genetic engineering, beneficial bacteria play a significant economic role. Ruminococcus species aid in the digestion of cellulose by secreting the enzyme cellulase, whereas Lactobacillus is utilized to manufacture curd and cheese. Escherichia coli is used to prepare vitamin K and riboflavin. Humans are vulnerable to several illnesses caused by pathogenic bacteria, such as tuberculosis (TB) caused by the mold Mycobacterium tuberculosis and food poisoning caused by saprotrophic microorganisms that target and break down tissues . In order to comprehend the behavior of various bacterial species, it is crucial to categorize them using machine learning approaches. Many researchers have used machine learning techniques to examine different kinds of bacteria.

1) In 2011, Rulaningtyas et al. implemented NN trained with back propagation approach for the classification of tuberculosis bacteria. The classification was done using NN with fine-tuned hyper-parameter values i.e. momentum $= 0.9$, learning rate $= 0.5$, mean square error $= 0.00036$ and number of hidden layers $= 20$. The dataset used in the experiment consisted of 100 samples of tuberculosis binary images out of which 75images were used for training process and 25 images for testing process. The given approach delivered and accuracy of 80%.

2) An automated classification method was introduced by Khutlanget al. (2010) in order to identify Mycobacterium tuberculosis images of ZN-stained sputum smears. The author used KNN, PNN and SVM classifier for classification of objects. The dataset used in the experiment consisted of total 11,259 instances of bacteria, out of which, 6901 instances from 11 subjects were used for training, which further consisted of 4999 instances of bacilli and 1902 instances of non-bacilli. For testing purpose, a total of 468

instances of bacteria obtained from 8 subjects were used, out of which, 1838 instances were of bacilli and 2520 instances were of non-bacilli. The given approach achieved the sensitivity and specificity values of 95%.

3) Akova et al. in 2010 proposed a ML based method to classify bacterial serovars. A total of 28 subclasses from 5 different bacterial species (*E. coli, staphylococcus, salmonella, vibrio and Listeria*) were used in the given experiment. The dataset consisted of 2054 random samples of the bacteria which were then split into 80% and 20% for training and testing purpose respectively. Initially, the textural features were obtained from the samples which were then fed to the Support vector data description (SVDD) and Bayesian classifier. The Bayesian classifier was not specific to bacteria classification problem, thus the results for the Bayesian classifier were validated on the benchmark letter recognition dataset from the UGI repository. The experiment was able to achieve an accuracy of 82%.

4) Govinda et al. introduced a technique for the classification of tuberculosis bacilli bacteria using ZN images. The given approach used SVM with library tools (LIBSVM) for classification purpose. The experiment was carried out using two microscopic image sets: the first one was obtained from Public health Image library (PHIL), an open source for microscopic images and second was obtained from Global Hospital and health city, Chennai. These image sets were then used to extract the ZN stained images of 34 tuberculosis positive and 16 negative patients. The proposed method achieved an accuracy of 90.89% with sensitivity of 72.89%.

5) Chayadevi et al. employed both statistical and NN approach for extracting bacterial clusters and counting different bacterial species from digital microscopic images. The proposed technique involved thresholding method and binarization; followed by segmentation and feature extraction. Finally, K-means clustering and self-organizing maps (SOM) were used for clustering and counting purpose. Authors used primary dataset collected from the hospital that consisted of 320 digital images of bacteria. The results obtained by using the proposed system were compared with the manual count taken by the doctors, wherein the proposed system proved to be more accurate than human visual counting.
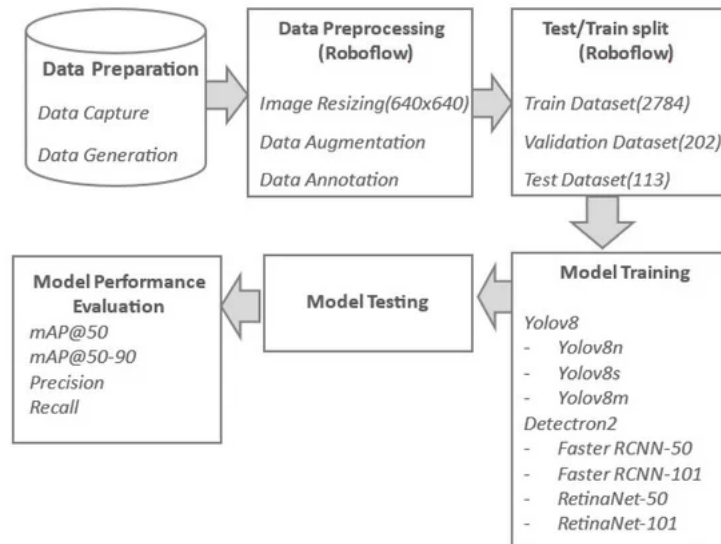
Chapter 4

Methodology

Fig. 1 The step by step process involved model training

4.1. *Dataset and Preprocessing*

The collection and processing of data are crucial steps in any research. The section explains how the dataset was gathered, followed by some processing techniques to prepare it. The images were distributed into three folders: 429 images for training, 40 images for validation, and 20 images were kept aside for final testing.

4.2 Initial Data Collection

Several sample papers were studied to find out datasets but not much data was available on the internet so 489 sample images from a private repository were sent via email.

To solve small dataset issue, various methods like data augmentation and transfer learning have also been employed by the researchers. Although these methods are also insufficient and can carry other related problems. Like in transfer learning one cannot provide suitable convolutional filters specific to the task and in data augmentation unknown pattern and data

cannot be managed. It is very difficult to create large datasets, because the most of the technologies are patent and lack the thorough involvement of microbiologists. The optimal way to improve the efficiency of deep learning model is to work in collaboration with microbiologists, to design and annotate the labeled benchmark datasets.

In total, around 30% percent of the dataset consists of generated data. An argument against using this type of generation could be that this could make the model too focused to detect a specific type of bacteria. Still, the model predicts the bacteria directly as classes. The model learns the class of the annotated bullet hole mainly based on the context of where it is located and not by the way it looks.

4.3. Data Annotation

A total of 489 sample images were annotated using Roboflow API. Technically, only the bacteria classes are necessary to calculate the score. The last column of the table represents how many times each class has been annotated in the dataset

. 4.4. Data Augmentation

Before the annotated dataset was used for model training, images were resized to 640 × 640 in the coco format. The reason for resizing is to adjust images for the pre-trained Detectron models as they require the images in that shape. Secondly, the dataset was expanded to increase its size and avoid overfitting. The size of the dataset is developed by creating multiple copies of each original image. These numerous copies are made by applying operations on the original image. This dataset expansion technique is known as data augmentation. shows all the implemented augmentations using various python libraries.

4.5. Models

YOLO

YOLO belongs to the single-shot detection category, and its first version was published in 2015. It carries out object detection by splitting the image into a k × k grid of equal dimensions. Each grid is responsible for detecting the object if its center falls inside the grid. A grid can predict a few bounding boxes with a specific confidence value. Each predicted bounding box comprises five values, i.e., x and y values for the center of the bounding box, width, height, and confidence value. To choose the most representative bounding box in the grid and remove the excess boundary boxes, YOLO uses the intersection over union (IoU) and non-max suppression (NMS) methods, respectively. YOLOv8 is the latest version of the

YOLO algorithm and was released by Ultralytics in July 2023. The entire architecture of YOLOv8 can be split into two main parts, the backbone and the head.

Detectron2

Detectron was released by Facebook AI Research on 9 October 2019, as an upgrade to the original detectron2 framework. This framework is primarily designed for object detection and instance segmentation tasks. The framework is built on the PyTorch deep learning library, enabling integration with other neural network architectures and facilitating experimentation. Detectron2 consists of four main components: the backbone, the neck, the region proposal network (RPN), and the head shows the architecture of detectron2. The backbone is responsible for feature extraction from the input image, using various architectures such as ResNet, ResNeXt, and MobileNet. These architectures are often pre-trained on large-scale image datasets like ImageNet. The pretraining process involves training a network on many labeled images from ImageNet to learn general visual representations. ImageNet is an extensive dataset that leverages the hierarchical organization provided by WordNet, a lexical database. The backbone network consists of several convolutional layers organized hierarchically. These layers gradually down-sample the spatial dimensions of the feature maps while increasing the number of channels.

4.6 Evaluation Metrics

After training and validating the models, the performance of the trained models is measured by running the model on the test dataset. Choosing the right metrics for assessing the performance of object detection models can be challenging. Mostly, the models use precision, recall, and mean average precision (mAP) as metrics. These metrics are defined as follows.

Precision and Recall

Precision is the measurement of the correct predictions made by the model, indicating the model's reliability for truthful predictions. However, recall refers to the number of measurements the model correctly predicted as relevant to the pertinent total predictions. Below equations present the expressions used to calculate these scores.

$$\text{Precision (P)} = TP/TP + FP$$

(1)

$$\text{Recall (R)} = \text{TP/TP} + \text{FN}$$

$$(2)$$

where TP refers to true positives (when the algorithm correctly detects a bullet hole with the bounding box), FP represents false positives (when the bounding box is generated on a location without having any bullet hole), and FN corresponds to false negatives (when a bullet hole is undetected). The IoU between the ground truth and the generated bounding box is calculated using Equation (4). In each image, for every object, if the value of IoU is above the predetermined threshold (mostly 0.5), then that is regarded as a TP; otherwise, the result is an FP. A trained model generates a TP by using the coordinates of the bounding box and the confidence score (the confidence for each detection made by the model).

Average Precision and Mean Average Precision

Average precision (AP) is the area under the precision–recall curve and is calculated as shown below.

$$\text{AP}=\sum n-1k=0[\text{Recall(k)}-\text{Recall(k+1)}]\times\text{Precisions(k)}$$

$$(3)$$

$$\text{IoU} = (\text{Object} \cap \text{Detected box})/(\text{Object} \cup \text{Detected box})$$

$$(4)$$

The value of AP is between 0 and 1 to summarize the different precision values into a single value representing the average of all precisions. Moreover, mean average precision (mAP) is calculated by taking the average of AP values for all classes and is shown below.

$$\text{mAP}=1n\sum ni=1[APi]$$

$$(5)$$

Primarily, two different threshold values are used by object detection models, i.e., mAP50 (mean of AP with confidence values between 0 and 0.50) and mAP50-90 (the mean of AP with confidence scores between 0.50 and 0.95). For object detection models, precision, recall, mAP50, and mAP50-95 are considered the most common metrics [32]. The detectron2 framework generates some other metrics, as listed below:

- AP75: This metric calculates the average precision at 75% overlap. It measures the precision at different recall values, considering a detection as TP if the IoU overlap is at least 75%.

- APl: The "average precision (large)" calculates the precision at different recall values for objects with large sizes (area > 96^2). It focuses on evaluating the performance of detecting larger objects accurately.

- APm: This metric is the average precision (medium) and measures the precision at different recall values for objects with medium sizes (32^2 > area > 96^2). It is used to assess the performance of detecting objects of moderate dimensions.

- APs: The "average precision (small)" calculates the precision at different recall values for objects with small sizes (area < 32^2). It evaluates the model's ability to detect smaller objects accurately

4.7 Steps performed for this project

1) The first step was to annotate the dataset and it was done using roboflow.

2) The second step was to download this in the coco.yaml format which is the only format supported by detectron2.

3) The next step was to install the detectron2 libraries and the required dependencies.

4) Now, import detectron2 and its modules and some common libraries that would be requires in this process

5) First, we have to define the complete configuration of the object detection model. We imported the 'get_cfg' function from the detectron2.config module, we will be using it now. I have chosen the Coco Instance segmentation configuration (YAML file). There are other options available too. You also have to set the model's threshold score (usually set between 0.4 to 0.6).

6) Next, use the Visualizer class to see how detection has been performed. The visualize class has a function to draw instance predictions.

7) Our function will input the image directory/folder path as input. The JSON files are is then opened and loaded. We enumerate through records of JSON files, get the image path. Each image is read from the path, and its height, weight, file name, and image ID are stored in a dictionary 'record' Next, we read through the annotations, and store bounding box details in another dictionary 'obj'. At the end of each loop, the record is appended to a list called "dataset_dicts".

8) Similarly, the bounding box dictionaries are also appended to a list "objs". This list will in turn be assigned as the value against the "annotations" key in the record dictionary. Each of these record dictionaries is then appended to a final list which will be returned.

9) The next step is to register these training and validation datasets. For registering a dataset, you have to use the DatasetCatalog.register and the MetadataCatalog method.

10) Now that we have registered the datasets we will train the model, start by importing DefaultTrainer from the engine module of Detectron. Define the dataset and other parameters like no of workers, batch size, no of classes (8 in this case). We initialize the model with pre-trained weights and train further. The max iterations parameter will vary depending upon the size of the dataset and the complexity of the task.

11) Then we load the model into a python file, let's call it app.py and then we import the flask module which would act as a backend to store the results of the module and then we write the python code to load the image and inference it with the model. We link this file app.py with an html template which would be used as a front end for this project.

The metrics for our data are as follows :

Precision: 66.1%

mAP: 74.8%

Recall:74.8%

Fig. 2 Index.html front end

Chapter 5

Results
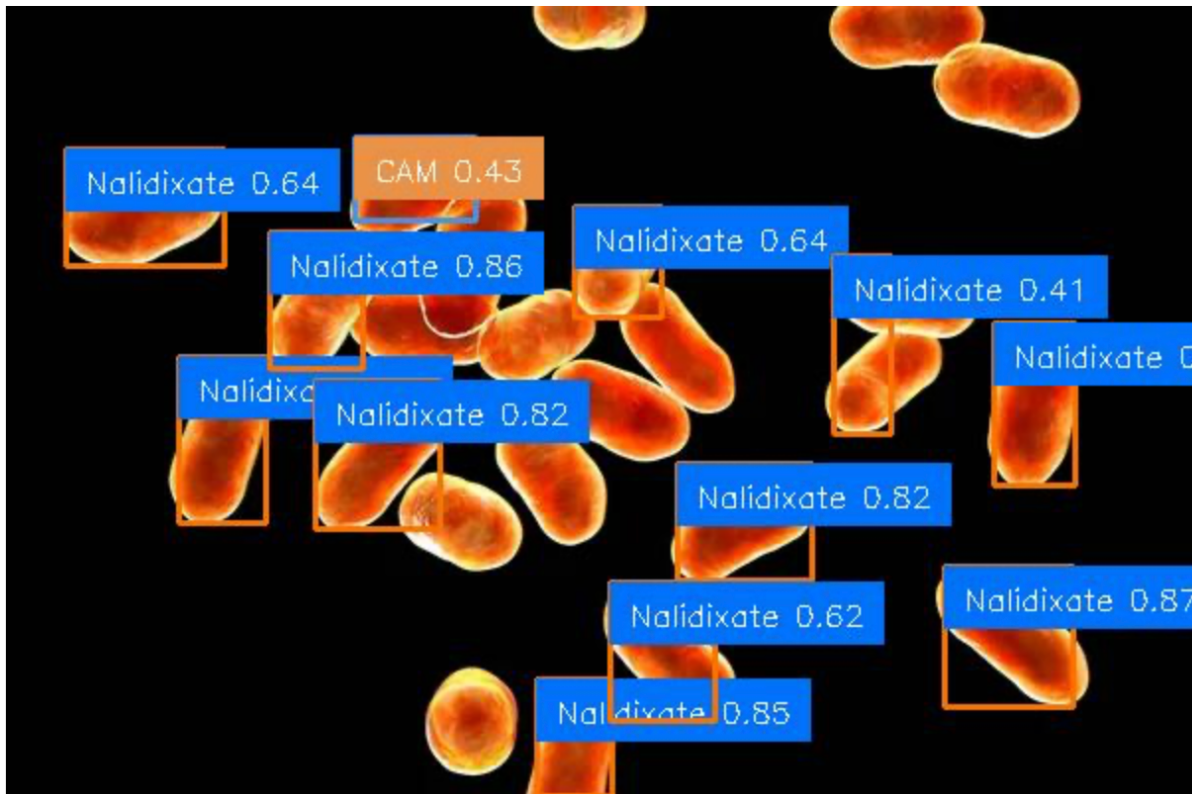

Fig.3   Visualised Results

Chapter 6

Conclusion and Future Scope

An overview of the different ML techniques used in the bacterial classification is provided . ML techniques have been extensively applied in the field of bacterial classification. The field of microbiology has improved with the application of these techniques. ML techniques help the microbiologist in many aspects i.e. in the identification, classification of bacteria and also the over-all automation of these processes. The aim of the review is to explore different ML based methodologies used by various researchers in the bacterial image classification. Different imaging and data modalities have been explored in the domain of bacterial classification. The data collected in this paper highlights the efforts put by the researchers in their studies. However it is difficult to compare their methods and performances due to the difference in the dataset and imaging modalities used by each of them. It can be noted that the variation in their performances may also be attributed to the different approaches and methods used in their research work. It has been observed that ML techniques proved to be a better approach, by giving better accuracy and precision values during the classification of bacteria. DL methods have also been used extensively for the classification of bacteria. In DL some state-of-the-art algorithms are also used i.e. deep belief network, CNN and LSTM for classification purpose. ML has been successfully used in the classification of bacteria. n future, the researchers can combine DL techniques with other ML approaches for getting better results. Researchers can explore other DL approach like recursive neural networks, auto encoders etc. in microbiology field. The dataset can also be enhanced by using different species of bacteria and then classify that data with different ML techniques or combination of such techniques. There is a great potential in future for ML techniques in the field of microbiology. Researchers continue to explore other domains like informatics, medicine, entomology and biology where ML techniques can be used.

Chapter 7

References

[1]. Indira B, Valarmathi K. A perspective of the machine learning approach for the packet classification in the software defined network. *Intell Autom Soft Comput.* 2020;26(4):795–805. doi: 10.32604/iasc.2020.010114..

[2]. Kalaivaani P, Thangarajan DR. Enhancing the classification accuracy in sentiment analysis with computational intelligence using joint sentiment topic detection with medlda. *Intell Autom Soft Comput.*

[3]. "Bacteria", https://en.wikipedia.org/wiki/Bacteria, 23 Nov 2020. IEEE.

[4] Wang W, Yang Y, Li J, et al. Woodland labeling in Chenzhou, China, via deep learning approach. *Int J Comput Intell Syst.* 2020;13(1):1393–1403. doi: 10.2991/ijcis.d.200910.001.

[5]. . Li C, Wang K, Xu N. A survey for the application of content-based microscopic image analysis in microorganism classification domains. *Artif Intell Rev.* 2019 doi: 10.1007/s10462-017-9572-4.

[6]. Panicker RO, Kalmady KS, Rajan J, Sabu MK. Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. *Biocybern Biomed Eng.* 2018 doi: 10.1016/j.bbe.2018.05.007.

[7].Yasuki Kakishita;Hideharu Hattori;Anirban Ray;Akiko Hisada;Yusuke Ominami;Jean-Pierre Baudoin;Jacques Yaacoub Bou Khalil;Didier Raoult(2004, January)

[8]. Holmberg M, Gustafsson F, Gunnar Hornsten E, Winquist F, Nilsson LE, Ljung L, Lundstro I. Bacteria classification based on feature extraction from sensor data. *Biotechnol Techn.* 1998;12(4):319–324. doi: 10.1023/A:1008862617082.

Chapter 8

Plagiarism Report

Chapter 3

Literature review

Intensive research was done and articles were searched and gathered from famous sources like IEEE Explore, Google Scholar, Research gate. One subset of AI that has excelled in computer science is machine learning. After being trained on a set of data, machine learning algorithms may be used to forecast the results of fresh data. Researchers in microbiology analyze digitized tiny bacterial photos using machine learning techniques. Bacteriology is the study of bacteria; researchers examine the bacteria using their genetic makeup and physical characteristics. In several fields, including food processing, biotechnology, fiber retting, pest management, and genetic engineering, beneficial bacteria play a significant economic role. Ruminococcus species aid in the digestion of cellulose by secreting the enzyme cellulase, whereas Lactobacillus is utilized to manufacture curd and cheese. Escherichia coli is used to prepare vitamin K and riboflavin. Humans are vulnerable to several illnesses caused by pathogenic bacteria, such as tuberculosis (TB) caused by the mold Mycobacterium tuberculosis and food poisoning caused by saprotrophic microorganisms that target and break down tissues . In order to comprehend the behavior of various bacterial species, it is crucial to categorize them using machine learning approaches. Many researchers have used machine learning techniques to examine different kinds of bacteria.