Arab Academy for Science, Technology and Maritime Transport

College of Computing and Information Technology

Computer Science Department

B. Sc. Final Year Project

Smart Microscope for Evaluating Water Quality

Presented By:

Sara Omar Abdelaziz

Nehal Mostafa Kamel

Ola Mohamed Mahmoud

Aya Ehab Youssef

Radwa Abd Elhakeem

Nada Mohamed Mostafa

Supervised By:

Prof. Nashwa El-bendary

JULY – 2022

DECLARATION

I hereby certify that this report, which I now submit for assessment on the programme of study leading to the award of Bachelor of Science in <*insert title of degree for which registered*>, is all my own work and contains no Plagiarism. By submitting this report, I agree to the following terms:

Any text, diagrams or other material copied from other sources (including, but not limited to, books, journals, and the internet) have been clearly acknowledged and cited followed by the reference number used; either in the text or in a footnote/endnote. The details of the used references that are listed at the end of the report are conforming to the referencing style dictated by the final year project template and are, to my knowledge, accurate and complete.

I have read the sections on referencing and plagiarism in the final year project template. I understand that plagiarism can lead to a reduced or fail grade, in serious cases, for the Graduation Project course.

Student Name: Sara Omar Abdelaziz
Registration Number: 18108672
Signed: ______________________
Date:

Student Name: Nehal Mostafa Kamel
Registration Number: 18108704
Signed: ______________________
Date:

Student Name: Ola Mohamed Mahmoud
Registration Number: 18108550
Signed: ______________________
Date:

Student Name:Aya Ehab Youssef
Registration Number: 18108773
Signed: ______________________
Date:

Student Name: Radwa Abd Elhakeem
Registration Number: 18108670
Signed: ______________________
Date:

Student Name: Nada Mohamed Mostafa
Registration Number: 18108635
Signed: ______________________
Date:

STUDENTS CONTRIBUTION

Smart Microscope for Evaluating water Quality

Chapter	Title	Contributors
1	Introduction	Sara Omar Abdelaziz Nehal Mostafa Kamel
2	Literature Review	Ola Mohamed Mahmoud Aya Ehab Youssef
3	Proposed System	Radwa Abd Elhakeem Nada Mohamed Mostafa
4	Results and Discussion	Sara Omar Abdelaziz Nehal Mostafa Kamel Ola Mohamed Mahmoud Aya Ehab Youssef Radwa Abd Elhakeem Nada Mohamed Mostafa
5	Conclusions and Future Work	Sara Omar Abdelaziz Ola Mohamed Mahmoud Aya Ehab Youssef Nehal Mostafa Kamel

ACKNOWLEDGMENT

Our thanks to our supervisor PROF. NASHWA EL-BENDARY dean of Computer and Information Technology college for her kindness, love and patience in guiding us and supervising us to deliver this project. She has hugely inspired and encouraged us throughout the whole journey, it's a pleasure to be indebted to her for everything she has done.

Also, we would like to thank her because she had a huge role in teaching us Artificial Intelligence, Machine learning and Deep Learning courses that Benefited us hugely in this project.

To all of our family and friends thank you for helping us survive all along and encouraging us to give our best, and to our loving parents thank you for being our constant source of inspiration and motivation and for playing significant roles throughout our lives.

We would like to express our sincere gratitude to AAST for letting us be students here and for providing us the best education one could ever dream of, by improving our educational level as well as our skills; we are proud of being students in this university and we are indebted to such great professors for providing us necessary information.

And finally, to our teammates, thank you for your patience and enthusiasm, and for understanding, encouraging, supporting and pushing each other further than we thought we could go, we couldn't have done it without all of us.

ABSTRACT

Biomarkers have been used extensively to connect between external level of contaminant exposure ,internal level of tissue contaminant and early adverse effect in organisms . The health of a marine ecosystem can effectively be monitored by studying the levels of biomarkers in a representative species So, the water pollution level will be estimated from biomarkers(fish gills).
The idea of the project aims to design and develop a smart microscope system based on a deep learning model, which is one of the techniques of artificial intelligence techniques in order to determine water quality by biomarkers extracted from microscopic images of fish through that microscope without the need for a specialised technician (or in the absence or preoccupation of a specialised technician) and to identify the type and quantity of water pollution.
In this study, classification results are compared using different feature extraction algorithms that can extract various features of the fish gills image, successful GLCM, LBP, SIFT, SURF, HOG, and texture algorithms were selected as feature extraction algorithms.The features obtained from these methods are classified with SVM, KNN and random forest classifiers. The most successful feature extraction algorithm for fish gills images has been determined and the most successful classification algorithm has been determined. and also use a deep learning model (CNN, vgg16).
keywords: fish gills, deep learning model, water quality, digital microscope , HOG, LBP, GLCM, REC, SVM, KNN, CNN, vgg16

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

ACRONYM	Definition of Acronym
CART	Classification and Regression Tree
CNN	Convolutional Neural Network
DIC	Differential Interface Contrast
DoG	Difference of Gaussian
FITC	informed time-variant constraint
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
GLCM	Grey Level Co-occurrence Matrix
HOG	Histogram of Oriented Gradients
KNN	K-Nearest Neighbour
LBP	Local Binary Patterns
RBF	Radial Basis Function
SGD	Stochastic Gradient Descent
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UV	ultraviolet
VGG	Visual Geometry Group

Chapter One

INTRODUCTION

British poet W. H. Auden once noted, "Thousands have lived without love, but not one without water." Yet while we all know that water is crucial for life, we trash it anyways. Some 80 percent of the world’s wastewater is dumped—largely untreated—back into the environment, polluting rivers, lakes, and oceans. Today, 785 million people – 1 in 9 – lack access to safe water, Fig 1.1 shows that most of the countries have access to more than 85% or 85-70% improved water, more than 5 countries have 70-55% improved water, but on the other hand, more than 16 countries have less than 55% improved water, of which more than 10 countries have less than 55-40% and 6 countries have less than 40% improved water.

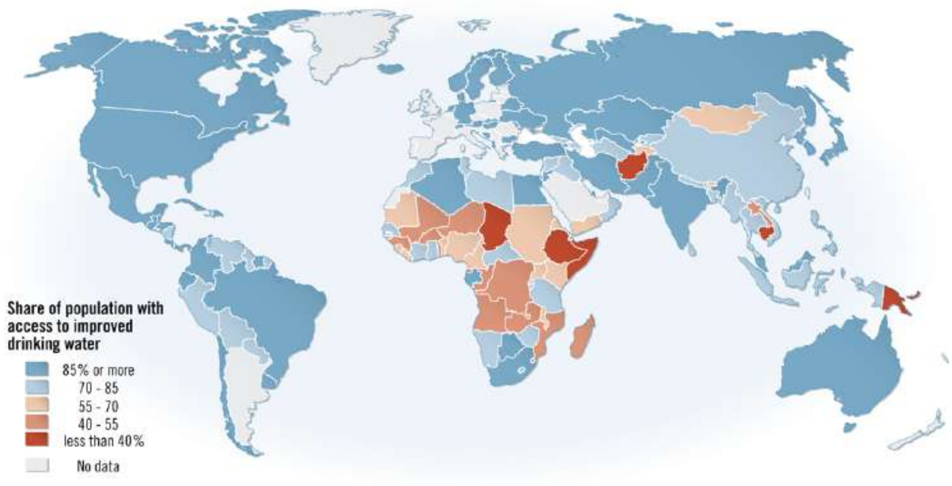


Figure 1.1 Share of population with access to improved drinking water [1]

This widespread problem of water pollution is jeopardising our health. Unsafe water kills more people each year than war and all other forms of violence combined. Meanwhile, our drinkable water sources are finite: Less than 1 percent of the earth’s freshwater is actually accessible to us. Without action, the challenges will only increase by 2050, when global demand for freshwater is expected to be one-third greater than it is now.

Water pollution is the contamination of water sources by substances which make the water unusable for drinking, cooking, cleaning, swimming, and other activities. These substances—often chemicals or microorganisms—contaminate a stream, river, lake, ocean, aquifer, or other body of water, degrading its quality and rendering it toxic to humans or the environment [2].

Water is uniquely vulnerable to pollution. Known as a "universal solvent," water is able to dissolve more substances than any other liquid on earth. It’s why water is so easily polluted.

Toxic substances from farms, towns, and factories readily dissolve into and mix with it, causing water pollution.

Pollution can directly kill or harm fish, or change the makeup of the fish's surroundings, killing off sources of food or causing plant or algae overgrowth that starves the fish of oxygen.

All animals breathe. How they breathe, however, varies.In order for us to breathe, we need air. So, what does that mean for animals that live in the ocean?

Fish need to take in oxygen and expel carbon dioxide in order to survive. But instead of lungs, they use gills. Gills are branching organs located on the sides of fish heads that have many, many small blood vessels called capillaries. As the fish opens its mouth, water runs over the gills, and blood in the capillaries picks up oxygen that’s dissolved in the water. Then the blood moves through the fish’s body to deliver oxygen, just like in humans. All bony fish also have a bony plate called an operculum, which opens and closes to protect the gills.

Fish aren’t the only undersea organisms to use gills, however. Mollusks and crustaceans also have gills that operate in the same way by pulling oxygen out of the water as it runs over the gills [3].

It's hard work to breathe underwater. The air we breathe has an oxygen concentration of 200,000 parts per million. Water, on the other hand, only has a concentration of 4 to 8 parts per million. That means fish need to run a lot of water over their gills to get the amount of oxygen they need to survive. Also, gills are very efficient in extracting the oxygen the fish needs.

Breathing gets even harder when the oxygen concentration decreases. You may have heard of hypoxic zones, also known as "dead zones," where low levels of oxygen make it impossible for animals to survive. Although they can occur naturally, hypoxic zones can also be created by human activity. When high-nutrient pollution, like farm runoff, enters the ocean, it can cause algal blooms that then die and decompose, causing low-oxygen zones. In 2019, runoff caused massive dead zones in the Gulf of Mexico.

Just like we need clean air to breathe, fish also need clean water to breathe. We can help gilled organisms thrive by keeping their ocean habitat free of pollutants.

The study was designed to investigate the influence of water pollution on gill apparatus or fish biomarkers in general.

Biomarkers are biological molecules found in blood, body fluids, or tissues that are considered a sign of normal or abnormal processes or diseases. Biomarkers have been widely used to establish a link between external contaminant exposure and internal contaminant levels in tissues, as well as early adverse effects in organisms. And also, it's used to see how well the body responds to a treatment for diseases or conditions. The health of a marine ecosystem can effectively be monitored by studying the levels of biomarkers in a representative species. So from fish biomarkers (fish gills, fish kidney, fish liver, etc.), we will determine the level of pollution in the water.

1.1 THE PROBLEM

Manual identification of water quality is a laborious process that relies heavily on experts. We measure water quality based on the biomarkers of fish or algae living in water bodies, which reflect the degree of water pollution. particularly in the context of drinking water. We decided to do this project to determine the quality of the water (high, medium, or low), as water quality testing is an important part of environmental monitoring. When water quality is poor, it affects not only aquatic life but the surrounding ecosystem as well. And also, by using our application software, we made it easier as it saves time and effort.

1.2 THE SOLUTION

This research aims to be able to maintain the ecological balance of the earth through controlling microorganisms that affect the safety of drinking water, thus saving the lives of all living beings. The finished product should be safe drinking water, free of microorganisms to a particular level.

use of a fish gill-based image classification technique as a biomarker to determine water quality (high, medium, or low).

1.3 THE OBJECTIVE

- Annotated databases of microscopic images dataset will be posted on the kaggle or mendeley data site for research databases.
- A Smart microscope system model based(Machine learning, Deep learning) to determine water quality by biomarkers extracted from fish gills microscopic images dataset.

1.4 MOTIVATION

Manual identification of water quality is a laborious process that relies heavily on experts. We measure water quality based on the biomarkers of fish or algae living in water bodies, which reflect the degree of water pollution. particularly in the context of drinking water. We decided to do this project to determine the water quality (high, medium or low) as water quality testing is an important part of environmental monitoring. When water quality is poor, it affects not only aquatic life but the surrounding ecosystem as well. And also, by using our application software we made it easier as it saves time and effort.

Chapter Two

BACKGROUND AND LITERATURE REVIEW

The use of biomarkers in aquatic sciences has long been established as an early warning, predictive and relatively low cost tool to indicate water pollution [4].

Heavy metals concentrations were determined in fish organs of Tilapia zilli and Clarias gariepinus from River Benue along Makurdi metropolis using atomic absorption spectrophotometer. The results indicated that Tilapia zilli gills contained the highest concentration (52.2%) of all the detected heavy metals, followed by the intestine (26.3%), while the muscle tissues appeared to be the least preferred site for the bioaccumulation of metals as the lowest metal concentration (21.5%) were detected in this tissue. Similarly, the Clarias gariepinus gills contained the highest concentration (40.3%) of all the detected heavy metals, followed by the intestine (31.6%), while the muscle tissue (28.1%) was the lowest. This study showed that the heavy metals were more concentrated in the gills of other parts of the fish in the water and showed that they were good biomarkers to monitor pollution in the river [5].

Combining Nomarski Differential Interface Contrast (DIC) and fluorescence microscopy using FITC and UV filters the system provides a reliable detection of micro-organisms with a considerable reduction in time, cost and subjectivity over the current labour intensive time consuming manual method [6] .

After conducting this empirical study, it was found that a combination approach in the segmentation of the algae works well. It was found that Otsu clustering along with the Kirsch filter is most suited for segmenting the microbes. Extensive Feature analysis for building machine learning dataset has been done. Correlation as a tool for feature selection was found to be an excellent choice for eliminating variables that are having multi-collinearity. It was found that at a correlation level of 0.85, the selected features provide stable and consistent results in terms of recall, precision, and accuracy for machine learning algorithms. Experimental evaluation and study of classification algorithms showed that CART Algorithm is best suited for this purpose [7].

Training a CNN from scratch requires a large amount of labeled data as well as high computational power. To overcome this challenge, the knowledge of a previously trained CNN model can be transferred to train new data with similar features. This technique is known as transfer learning. In this study, we use a pre-trained CNN ResNet-18 from PyTorch to classify bacteria division types by means of transfer learning. As described previously, we pre-processed the images applying data augmentation for the training and testing dataset. Due to the different sizes of the images, width between 35 and 4,295 pixels and height between 35 and 6,185 pixels, all images were resized to 128 ×128 pixels then randomly rotated, followed by random horizontal and vertical flips following the PyTorch recommendation. We train the model with the optimizer, Stochastic Gradient Descent (SGD).

The pre-trained model shows high test accuracy from the beginning and by epoch 5 (79.86 ±0.573 s) the accuracy is more or less stabilised at 99%. On the other hand, the non-pre-trained model stabilises at around epoch 12 (190.46±1.549 s). In the case of the pre-trained model, we could have run for fewer epochs, say 12, and still get essentially the same predictive

The behaviour of schools of zebrafish (Danio rerio) has been studied in acute toxicological environments. Behavioural features were extracted and a method for assessing water quality using a support vector machine (SVM) was developed. Behavioural parameters of fish were recorded and analysed within 1 hour in an environment with a 24 hour half-lethal concentration (LC (50)) of contaminants. The data were used to create a method for evaluating water quality in order to provide an early warning of toxicity. To enhance the efficiency and accuracy of the evaluation, a method combining SVM and Genetic Algorithm (GA) was used. The results showed that the average prediction accuracy of the method was more than 80% and that the time cost was acceptable. The method gave satisfactory results for a variety of mineral pollutants, demonstrating that this is an effective approach to water quality classification [8].

Chapter Three

Proposed system

In this chapter, we will know the implementation of the system to know the percentage of water quality for fish gills .

3.1: System Implementation

3.1.1 Hardware and software development tools, languages, etc

● Digital Microscope
● NVIDIA Jetson Nano Developer Kit (945-13450-0000-100)
● Microscopic slides
● Feature engineering (colour, shape and edge detection, skimage)
● ML algorithms(to apply classification based on hand-crafted feature (engineered features):
- Random forest algorithm - KNN - SVM
● Deep learning (knowledge) - VGG(Visual Geometry Group) - Convolutional Neural Network (CNN)
● Tensor-flow (google library / python library)

3.1.2 Experimental Analysis and Discussion

The dataset used for the experiments conducted was collected from Abbas farm, Abo-Hammad, Sharkia Governorate, Egypt. It consists of real sample images of fish gills exposed to copper and water pH in different histopathology stages. The collected dataset contains colored JPEG images as -- images and -- images were used as training and testing datasets, respectively. The training dataset is divided into three classes representing the different histopathology changes and water quality degrees as good, moderate, and bad water quality. Its content-based classification approach consists of three phases; namely Data Collection and preparation phase, feature extraction phase, classification phase, and decision phase as described in (Fig 3.2*)*.

(a) Good (b) Moderate (c) Bad

Figure 3.1. The labels of some samples of input image, which classify the water quality into three categories (Good, Moderate, and Bad)

3.2 Methodology

Phase 1: Data collection and preparation:

In phase 1, the dataset is being constructed through collecting images using a digital microscope, then the data is being prepared through assigning labels to the images.

1. Data collection: the data is collected from the fish's gills or liver. These fish exist in three environments:

a. Water with low pollution

b. Water with medium pollution

c. Water with high pollution

2. Data Preparation: on receiving the data with the support of the expert, the data is being labelled.

After the two phases of data collection and preparation, we applied an augmentation to the dataset to increase the number of images. We had 7 pictures. Now we have 270, 90 photos in each class (good, moderate, bad).

Generate images at random:

1. Rotate (20°)
2. Shift_width(5%)
3. Shift_height (5%)
4. Rescale the image by normalising it
5. Shear means cutting away part of the image (max 10%)
6. Zoom (10%)
7. horizontal flipping
8. Fill in missing pixels with the nearest filled value

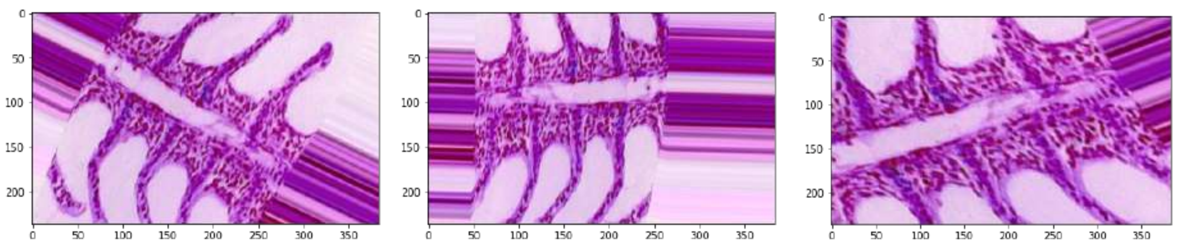


Figure 3.2. Samples of image augmentation

Phase 2: Feature extraction:

It is being worked with two methods:

1. Learn features by deep learning model

2. Feature engineering through colour, shape, and edge detection

Phase 3: Classification:

In it, training and validation is done to train the developed model to distinguish between batches of images based on the features or attributes extracted in the previous stage.

1- A classifier in machine learning is an algorithm that automatically orders or categories data into one or more of a set of "classes" (Decision Tree, Naïve Bayes classifier, KNN, SVM, Artificial Neural Networks)

2- SoftMax classifiers give you probabilities for each class label.

Phase 4: Decision:

Determining the extent of water pollution (high, low, medium) shows as (Figure 3.3).

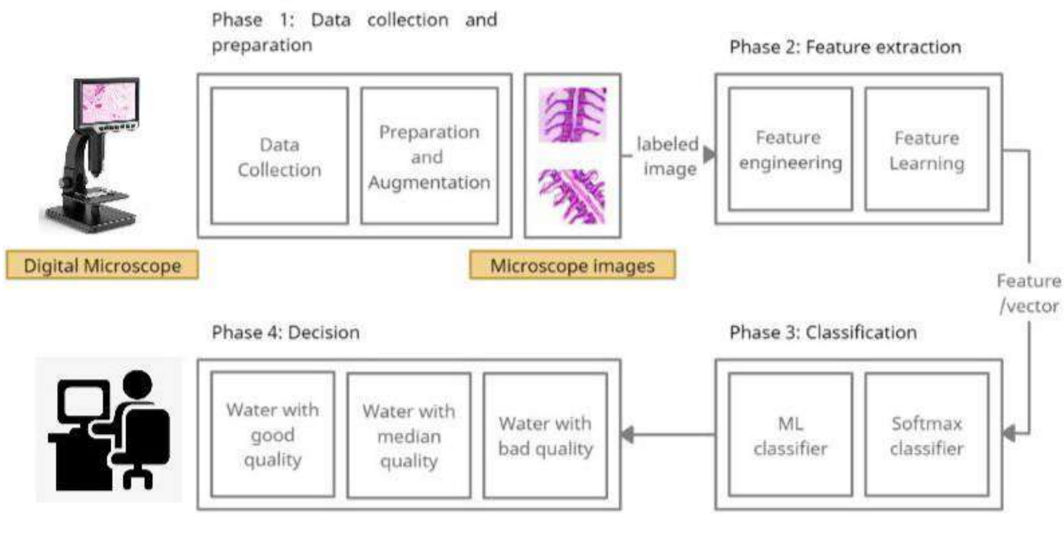


Figure 3.3 Methodology

3.3 System Design (Flowchart)

Figure 3.4 shows the whole process starting from augmentation till obtaining the Performance measures of different algorithms

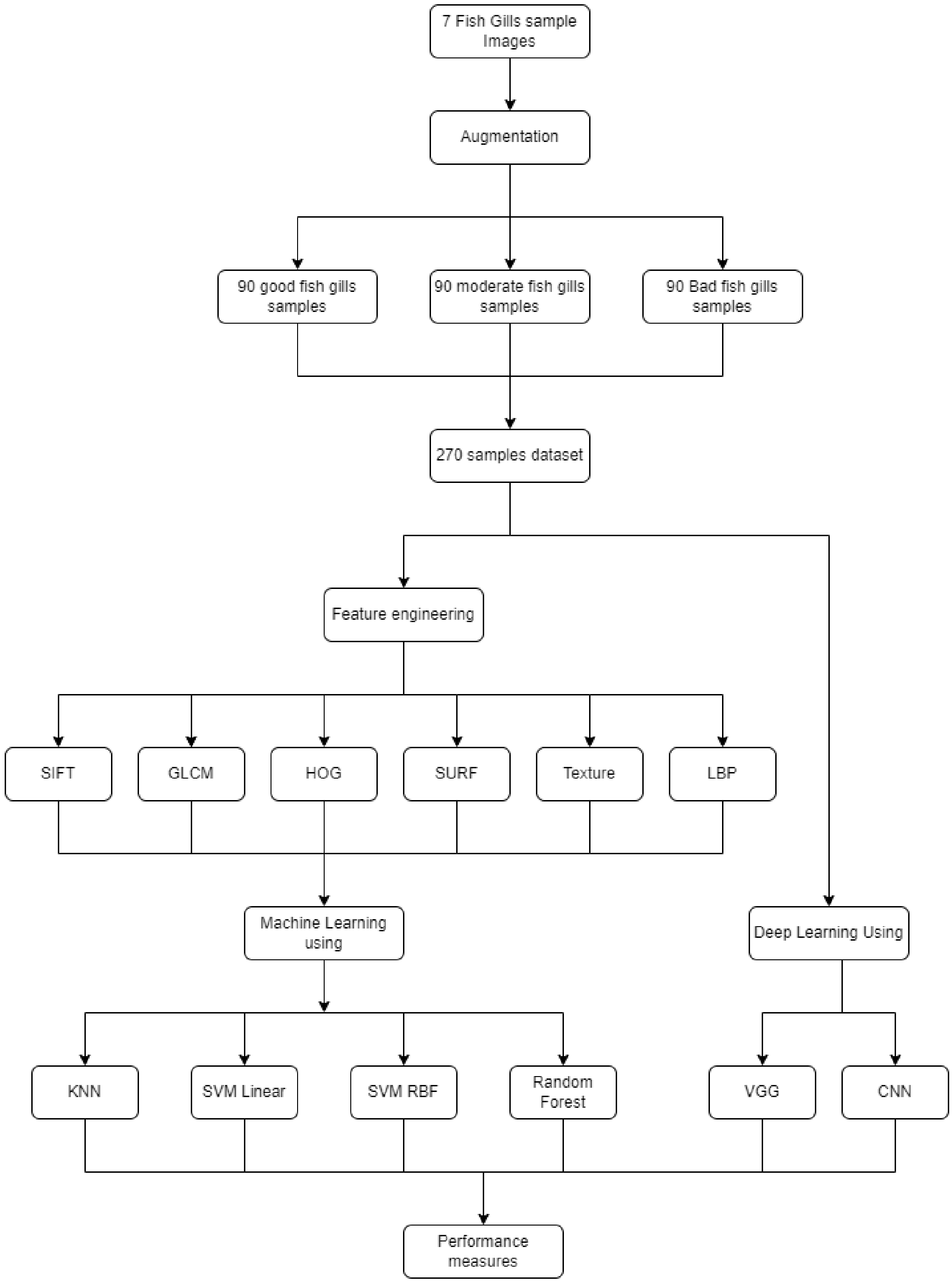


Figure 3.4 Flowchart

Chapter Four

RESULTS AND DISCUSSION

4.1 Introduction

This chapter contains detailed presentation and discussion of data analysis and the results of this study. The findings are presented under the following major headings:
Feature extraction, deep learning model (CNN, VGG-16), Feature engineering (colour, shape, ...), machine learning classifiers (KNN, SVM, ..)

4.2 Background information

4.2.1 Machine learning classifiers:

4.2.1.1 K-Nearest Neighbours (KNN)

KNN is an abbreviation of ‘K-Nearest Neighbours’ it’s one of the simplest supervised machine learning algorithms, it is used for both classification and regression predictive problems. KNN uses feature similarity to predict the values of new data points, by assigning the new datapoint a value based on how closely it matches the points in the training set [9].

4.2.1.2 Support vector Machine (SVM)

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, the classifier used is called Linear SVM classifier [10]

RBF: Radial Basis Function is a radial function where the reference point is not the origin. SVM with RBF kernel is capable of classifying data points separated with radial based shapes. Used for non-linearly separated data, which means the dataset cannot be classified by using a straight line.

4.2.1.3 Random Forest Algorithm

Random Forest is a supervised learning algorithm. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. It is considered a very good algorithm as it is used for both classification(which is based on majority voting) and regression(which is based on an average of all predictions) problems, which make up the majority of current machine learning systems. So the random forest algorithm is better than the decision tree (as the decision tree is very deep, which leads to overfitting). A random forest is a powerful modelling technique that is more powerful than a single decision tree. A random forest prevents overfitting by creating random subsets of features and building smaller trees using those subsets [11].

4.2.2 Performance Metrics

A confusion matrix is a table that is used to define the performance of a classification algorithm. The confusion matrix consists of four basic characteristics (numbers) that are used to define the measurement metrics of the classifier. These four numbers are:

1. TP (True Positive): is an output of the Confusion matrix, where the model predicts the positive class correctly.

2. TN (True Negative): is an output of the Confusion matrix, where the model predicts the negative class correctly.

3. FP (False Positive): is an output of the Confusion matrix, where the model incorrectly predicts the positive class as a negative class.

4. FN (False Negative): is an output of the Confusion matrix, where the model incorrectly predicts the negative class as a positive class [12].

Performance metrics of an algorithm are accuracy, precision, recall, Specificity, and F1 score, which are calculated on the basis of the above-stated TP, TN, FP, and FN.

$$\text{Accuracy} = TP + TN / TP + TN + FP + FN \quad (1)$$

$$\text{Precision} = TP / TP + FP \quad (2)$$

$$\text{Recall} = TP / TP + FN \quad (3)$$

$$\text{Specificity} = TN / TN + FP \quad (4)$$

$$F1 = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

4.3 Feature engineering

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.

The motivation is to use these extra features to improve the quality of results from a machine learning process, compared with supplying only the raw data to the machine learning process [13].

4.3.1 local binary patterns (LBP)

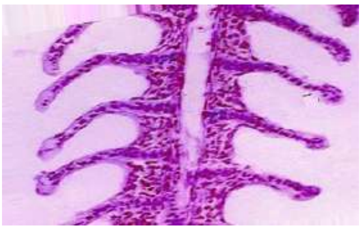
Local binary pattern feature extraction algorithm is a very useful algorithm that is resistant to light variations. We can simply describe the LBP process as follows; a window which has specified neighbourhood value is traversed over the image and a centre pixel label assignment is made. In this process, threshold is applied according to the pixel values adjacent to the centre pixel. Then, the LBP matrix is calculated according to the local neighbourhood values in the clockwise or counterclockwise direction. Thus, the statistical and structural model of the textural structure is calculated mathematically. The most important features of the LBP algorithm are resistant to grey level changes and computational simplicity which can be used in real-time applications . Equations 1 and 2 are used for labelling the pixels [14].

$$LBP_{P,R} = \sum_{P=0}^{P-1} S(G_P - G_c)^{2^P} \quad (6)$$

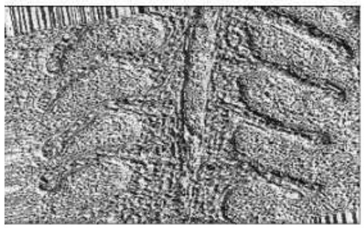
$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (7)$$

Where, G_r = centre pixel value

G_c = neighbour pixel value

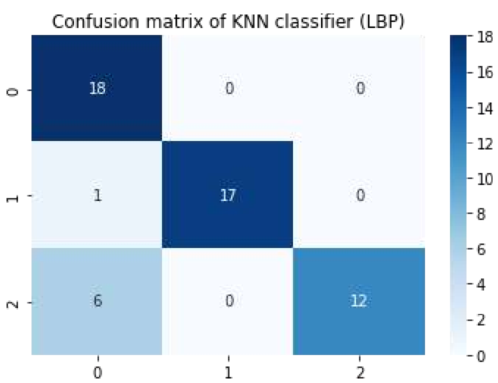


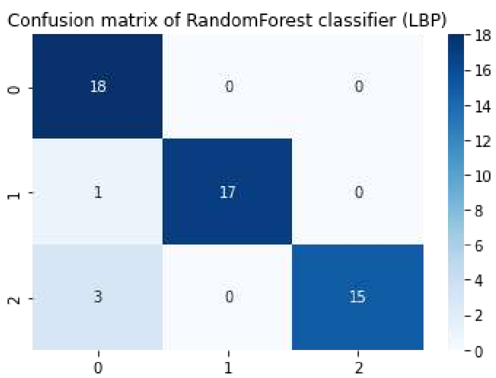
(a) **Before LBP**



(b) After LBP

LBP was applied to machine learning to determine the best accuracy





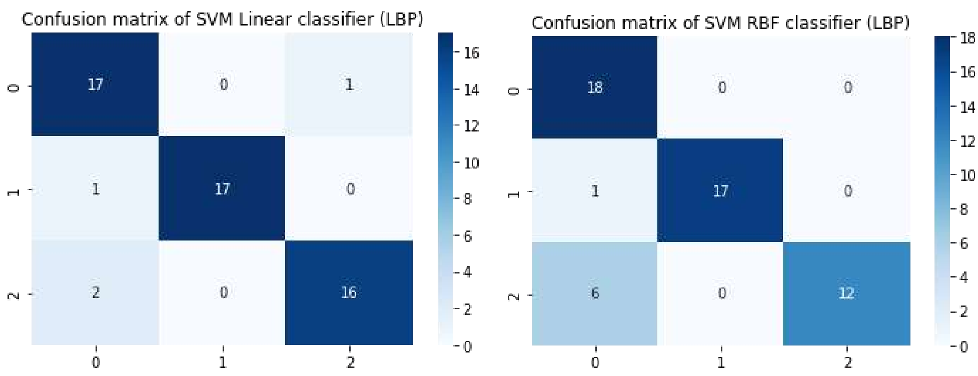


Figure 4.1 Confusion matrix(Testing) of LBP feature

Table 4.1 Comparison between Performance measurements of Testing (LBP)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.87	0.87	0.81	0.87	0.91
1	RFC	0.93	0.93	0.91	0.93	0.94
2	SVM(L)	0.93	0.93	0.94	0.93	0.93
3	SVM(RBF)	0.87	0.87	0.80	0.87	0.91

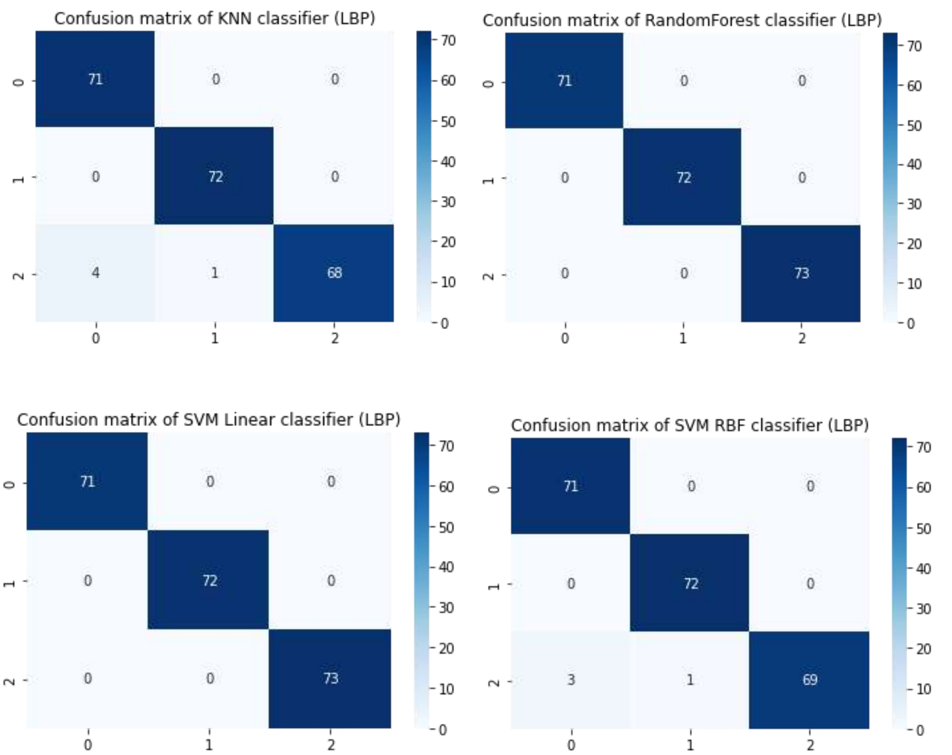


Figure 4.2 Confusion matrix(Training) of LBP feature

Table 4.2 Comparison between Performance measurements of Training (LBP)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.98	0.98	0.97	0.98	0.98
1	RFC	1.00	1.00	1.00	1.00	1.00
2	SVM(L)	1.00	1.00	1.00	1.00	1.00
3	SVM(RBF)	0.98	0.98	0.99	0.98	0.98

4.3.2 Colour feature (HOG)

hog is a shortcut for three words, histogram of gradient . It is used in computer vision and image processing for the purpose of object detection. The technique computes the occurrences of the gradient direction in the translated part of the image.[15]

To calculate the gradient of image:

Gradient is obtained by combining the magnitude and angle of the image. Given a block of 3 x 3 pixels, the first Gx and Gy for each pixel are calculated. The first Gx and Gy are calculated using this formulas for each pixel value:

$$G_x(r,c) = I(r,c+1) - I(r,c-1) \quad G_y(r,c) = I(r-1,) - I(r+1,c) \quad (8)$$

the gradient direction (orientation of edge normal) is given by :

$$Angle(\theta) = \left|tan^{-1}(G_y/G_x)\right| \quad (9)$$

the edge strength is given by the gradient magnitude:

$$Magnitude(\mu) = \sqrt{G_x^2 + G_y^2} \quad (10)$$

calculation steps:

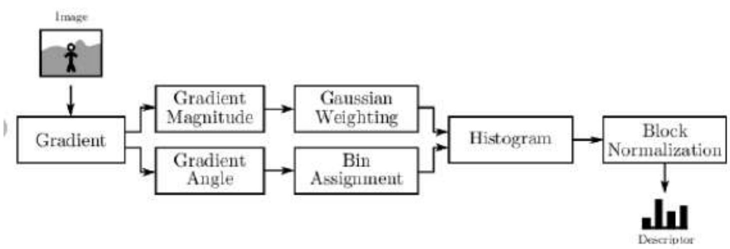


Figure 4.3 Calculation steps of HOG feature[16]

And then by applying colour feature(HOG) on the image:

(a) Before HOG

(b) After HOG

HOG was applied to machine learning to determine the best accuracy

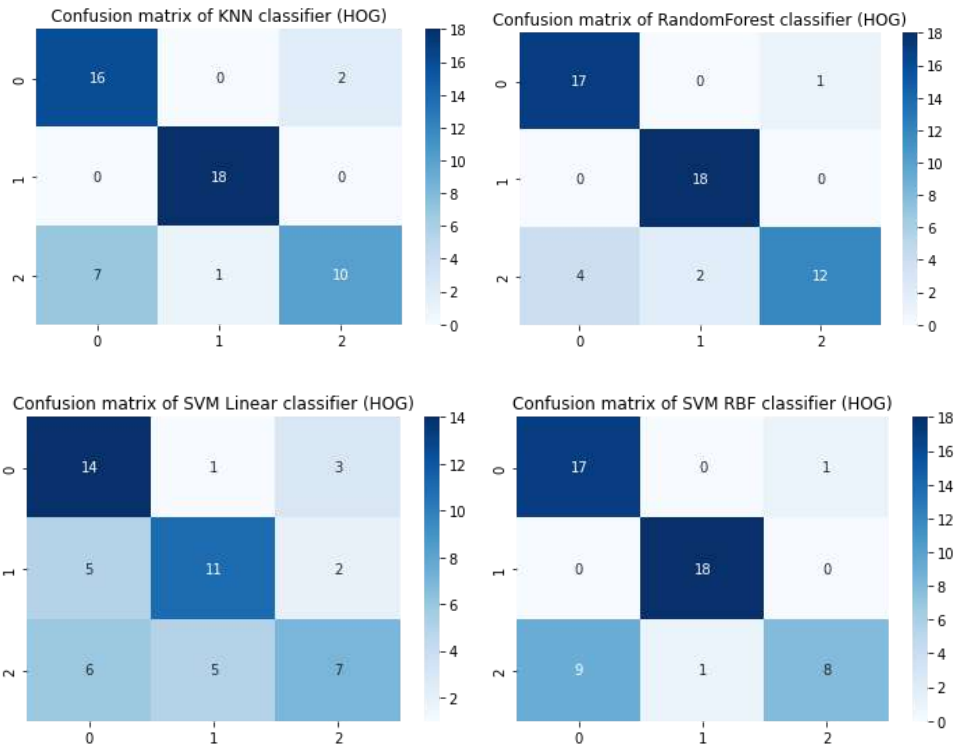


Figure 4.4 Confusion matrix (Testing) of HOG feature

Table 4.3 Comparison between Performance measurements of testing (HOG)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.81	0.81	0.81	0.81	0.83
1	RFC	0.87	0.87	0.89	0.86	0.88
2	SVM(L)	0.59	0.59	0.69	0.58	0.60
3	SVM(RBF)	0.80	0.80	0.75	0.78	0.83

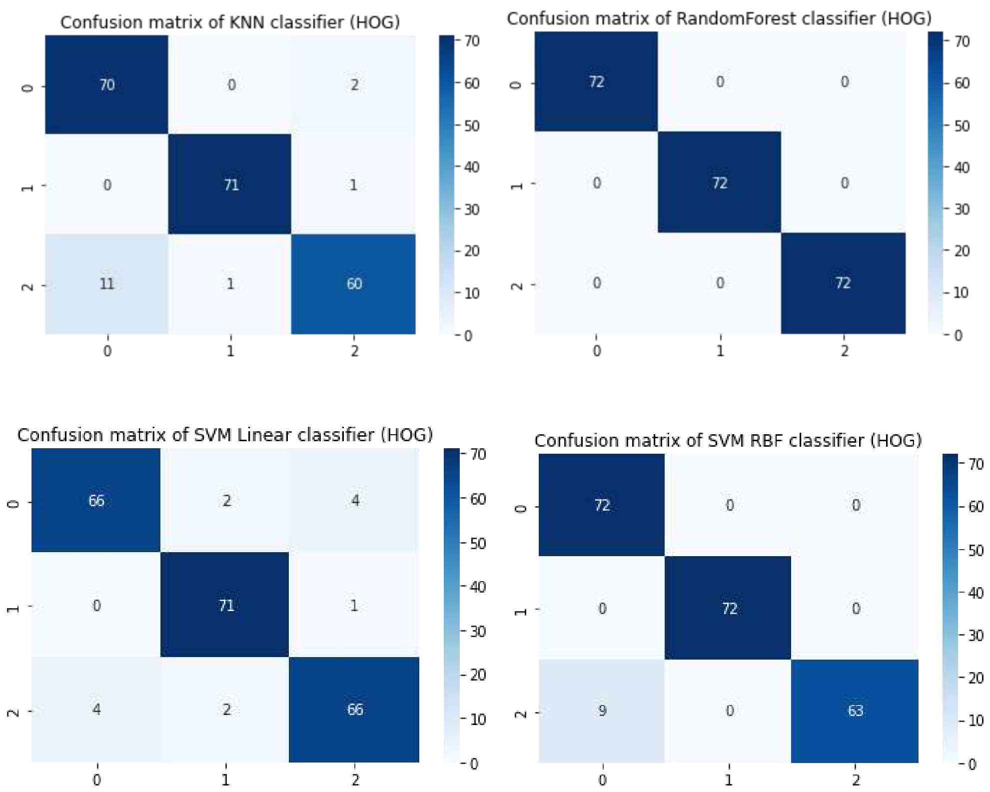


Figure 4.5 Confusion matrix (Training) of HOG feature

Table 4.4 Comparison between Performance measurements of training (HOG)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.93	0.93	0.92	0.93	0.93
1	RFC	1.00	1.00	1.00	1.00	1.00
2	SVM(L)	0.94	0.94	0.97	0.94	0.94
3	SVM(RBF)	0.96	0.96	1.00	0.96	0.96

4.3.3 Speeded-up robust feature (SURF)

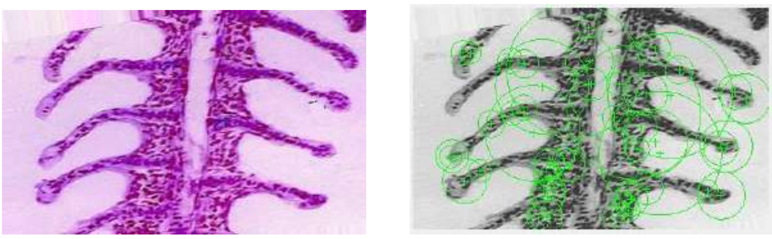
Speeded-up robust feature (SURF) is a robust feature detector which can get similar features from different objects used for various tasks such as classification and object recognition.

Speeded-up robust feature Have three steps: Interest point detection, Descriptor, and matching. To detect the Interest point SURF apply blob detector based on hessian matrix (2nd order spatial derivative) is defined as:

$$H(X,\sigma) = \begin{bmatrix} Lxx(X,\sigma) & Lxy(X,\sigma) \\ Lxy(X,\sigma) & Lyy(X,\sigma) \end{bmatrix} \qquad (11)$$

where $Lxx(X, \sigma)$ is the convolutional of the 2nd order spatial derivative of Gaussian with the image I(x , y) at the point X [17].

The descriptor describes the image features. It contains two steps: orientation assignment, and feature extraction. The direction of the point of interest points is measured to achieve rotational invariance, and this direction will be the dominant orientation. After that, the feature is extracted as a feature vector, which is matched with other image features.



(a) Before SURF **(b) After SURF**

SURF was applied to machine learning to determine the best accuracy, which is SVM(RBF).

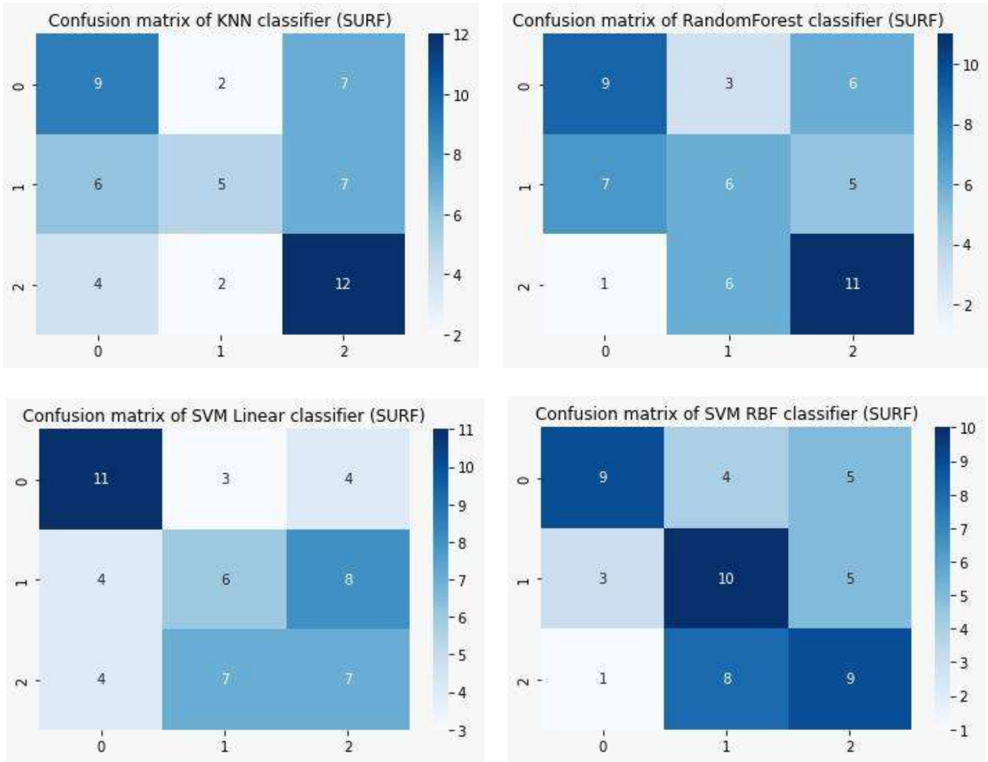


Figure 4.6 Confusion matrix of (Testing) SURF feature

Table 4.5 Comparison between Performance measurements of testing (SURF)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.48	0.48	0.72	0.47	0.50
1	RFC	0.48	0.48	0.78	0.48	0.48
2	SVM(L)	0.44	0.44	0.78	0.44	0.44
3	SVM(RBF)	0.52	0.52	0.89	0.54	0.52

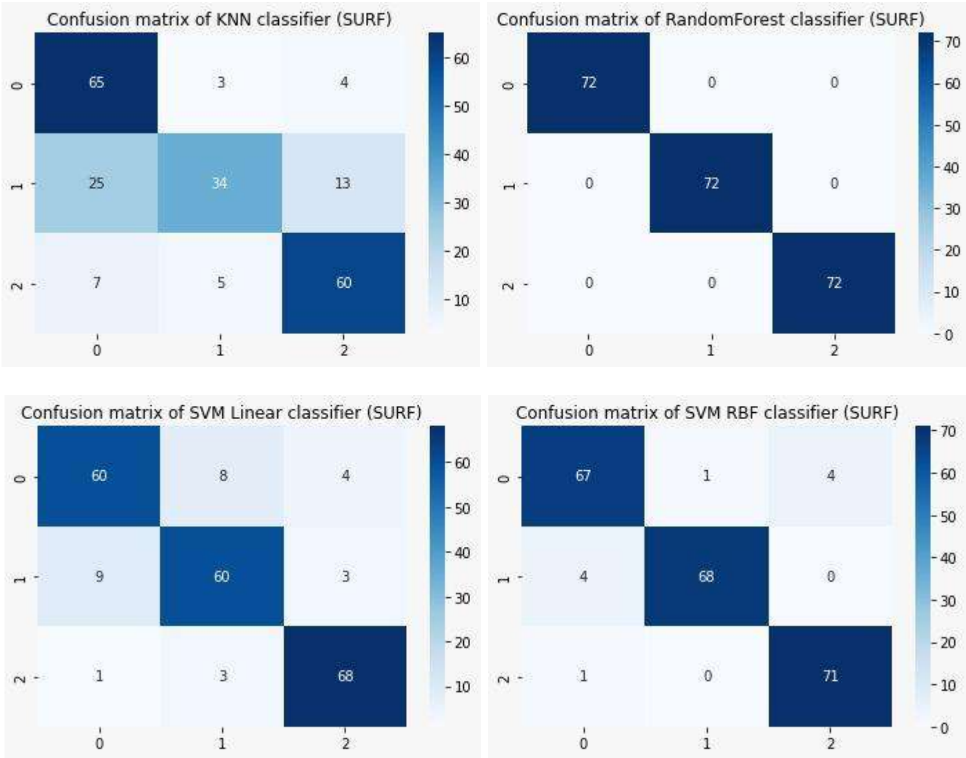


Figure 4.7 Confusion matrix of (Training) SURF feature

Table 4.6 Comparison between Performance measurements of training SURF

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.74	0.74	0.78	0.72	0.75
1	RFC	1.00	1.00	1.00	1.00	1.00
2	SVM(L)	0.87	0.87	0.93	0.87	0.87
3	SVM(RBF)	0.95	0.95	0.97	0.95	0.95

4.3.4 scale-invariant feature transform (SIFT)

SIFT is an abbreviation of scale-invariant feature transform, it's a computer vision algorithm that detects, describes, and matches local features in images.

SIFT feature is composed of several key points in the image with an orientation and corresponding descriptor of the area around the selected key points.

The algorithm finds the key points by searching through different image scales which is known as the Difference of Gaussian (DoG) pyramid.

The algorithm processes include Difference of gaussian (DoG) space generation, Key points Detection, and Feature Description[18].

1. **(DoG) space generation:**

   Image pyramid is a set of results that's obtained from the same image at different resolutions. This process has two steps:

   a) **smoothing the original image**

   The original image is passed through Gaussian smoothing filters with different kernel parameters to generate multiscale space, as Gaussian kernels are the only kernels that can generate them.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} exp(- \frac{(M/2-x)^2+(N/2-y)^2}{2\sigma^2}) \ (12)$$

   b/ gaussian kernel

   M denotes the length of the kernel and N denotes the height of the kernel, σ is the width of the Gaussian.

   then gaussian scale space is obtained from:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \ (13)$$

   c/ Gaussian scale space

   σ is the width of the Gaussian.

   b) **downsampling the processed image**

   Shrinks the image

   DoG space scale spaces can be obtained from:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \ (14)$$

   2 scales: σ and kσ

2. **key points detection**

key point is compared with 8 neighbours in the same layer and 9 neighbours in the upper and lower layers, to consider it as a key point it has to be the largest or smallest of the 27 points

3. **Feature description**

SIFT is the description of the gradient magnitude and gradient direction around the key points. In the first step, take the pixels of 16×16 centred on the key point.

Second allocate them to 4 × 4 blocks.

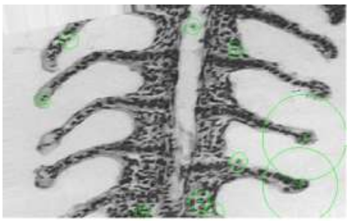
Third, divide 360° into eight directions on average, as 0°, 45°, and so on

Fourth, calculate the gradient magnitude of each direction in each block.

Last all the gradient magnitudes are expressed in one vector in order and the 4 × 4 × 8 = 128-dimensional SIFT feature is obtained.

After applying SIFT algorithm on the image this is the result we obtained:

(a) Before SIFT



(b) After SIFT

SIFT was applied to machine learning to determine the best accuracy

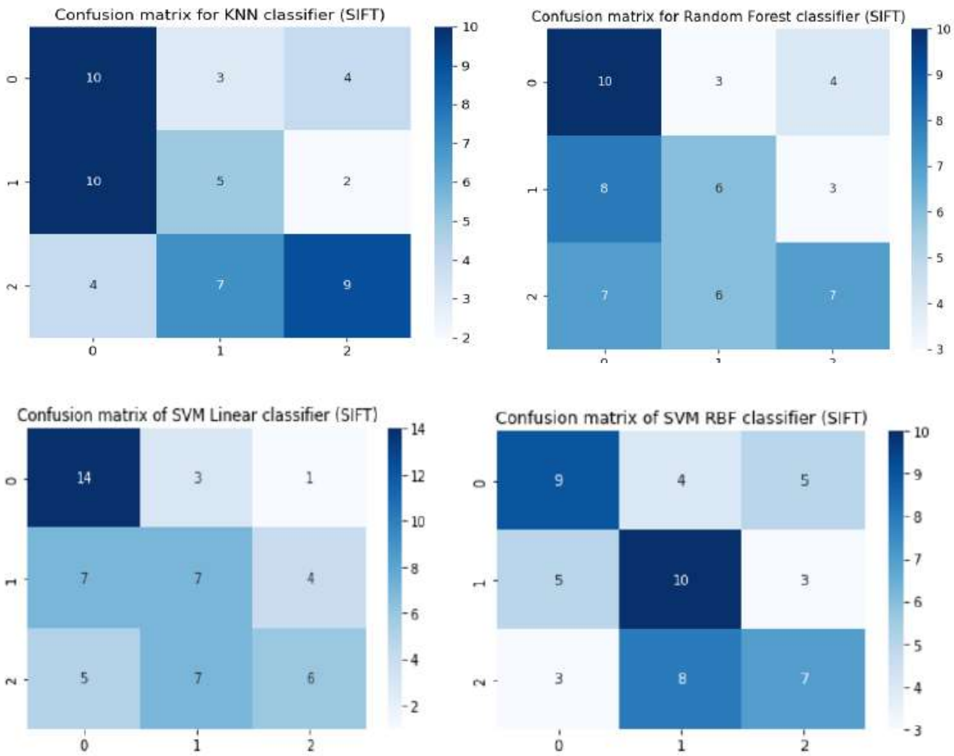


Figure 4.8 Confusion matrix of (Testing) SIFT feature

Table 4.7 Comparison between Performance measurements of testing(SIFT)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	**KNN**	**0.44**	**0.44**	**0.62**	**0.44**	**0.45**
1	**RFC**	**0.43**	**0.43**	**0.59**	**0.42**	**0.43**
2	**SVM(L)**	**0.50**	**0.50**	**0.67**	**0.48**	**0.50**
3	**SVM(RBF)**	**0.48**	**0.48**	**0.78**	**0.48**	**0.48**

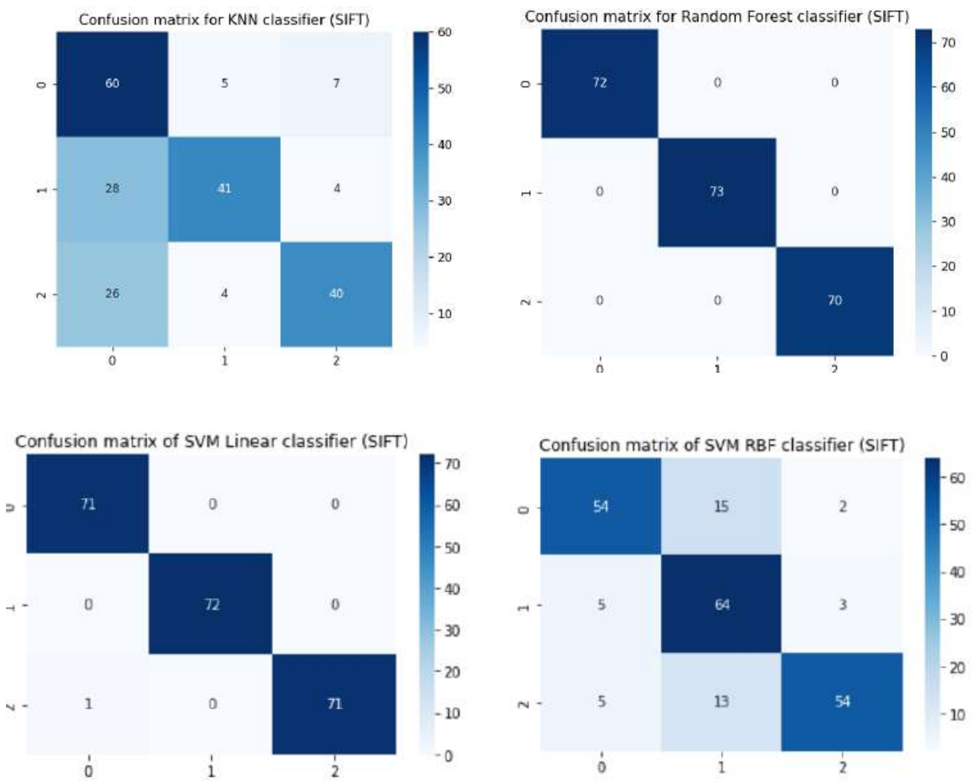


Figure 4.9 Confusion matrix(Training) of SIFT feature

Table 4.8 Comparison between Performance measurements of Training (SIFT)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.62	0.64	0.62	0.58	0.56
1	RFC	1.0	1.0	1	1.0	1.0
2	SVM(L)	1.0	1.0	0.99	1.0	1.0
3	SVM(RBF)	0.8	0.82	0.93	0.8	0.8

4.3.5 Texture

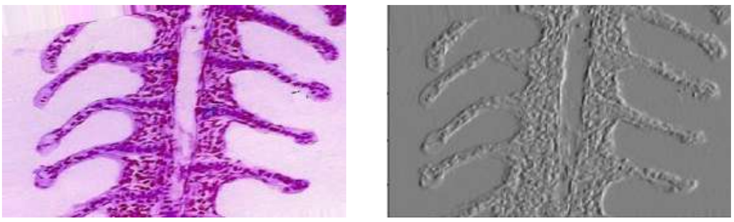
Texture is a key element of human visual perception and is used in many computer vision systems. For the eyes, distinguishing different textures is an easy task. Nevertheless, no precise definition of texture has been adopted yet. Some authors proposed to define it as a measure of coarseness, contrast, directionality, line-likeness, regularity, and roughness . The texture can also be seen as a similarity grouping in an image or as natural scenes containing semi-repetitive arrangements of pixels.Texture analysis is used in a very broad range of fields and applications, from texture classification (e.g., for remote sensing) to segmentation (e.g., in biomedical imaging), passing through image synthesis or pattern recognition (e.g., for image inpainting). For each of these image processing procedures, first, it is necessary to extract—from raw images—meaningful features that describe the texture properties[19].

$$ASM = \sum\sum P^2(i, j) \quad (15)$$

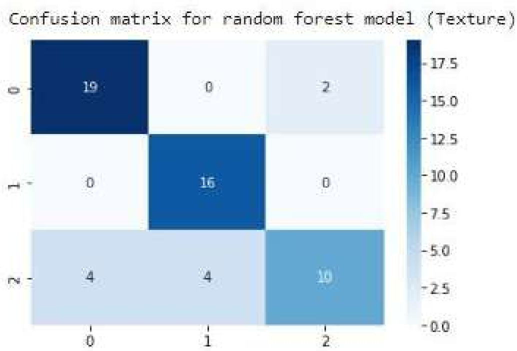
$$CON = \sum\sum(i - j)^2 P(i, j) \quad (16)$$

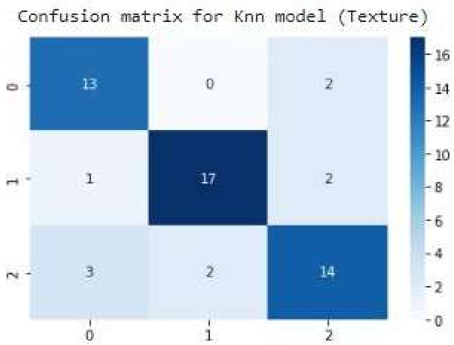
$$ENT = -\sum\sum(i, j) log[P(i, j)] \quad (17)$$

$$COR = \frac{\sum\sum ij\, p(i-j) - \mu_x\mu_y}{\sigma_x\sigma_y} \quad (18)\ [20]$$



(a) Before (b) After





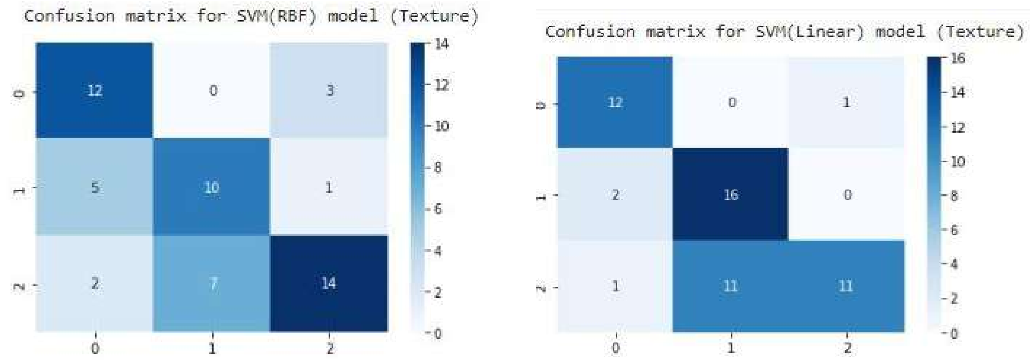


Figure 4.10 Confusion matrix (Testing) of Texture feature

Table 4.9 Comparison between Performance measurements of Testing (Texture)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.81	0.87	0.94	0.81	0.76
1	RFC	0.81	0.83	1.00	0.88	0.93
2	SVM(L)	1.00	1.00	1.00	1.00	1.00
3	SVM(RBF)	0.67	0.80	0.66	0.71	0.63

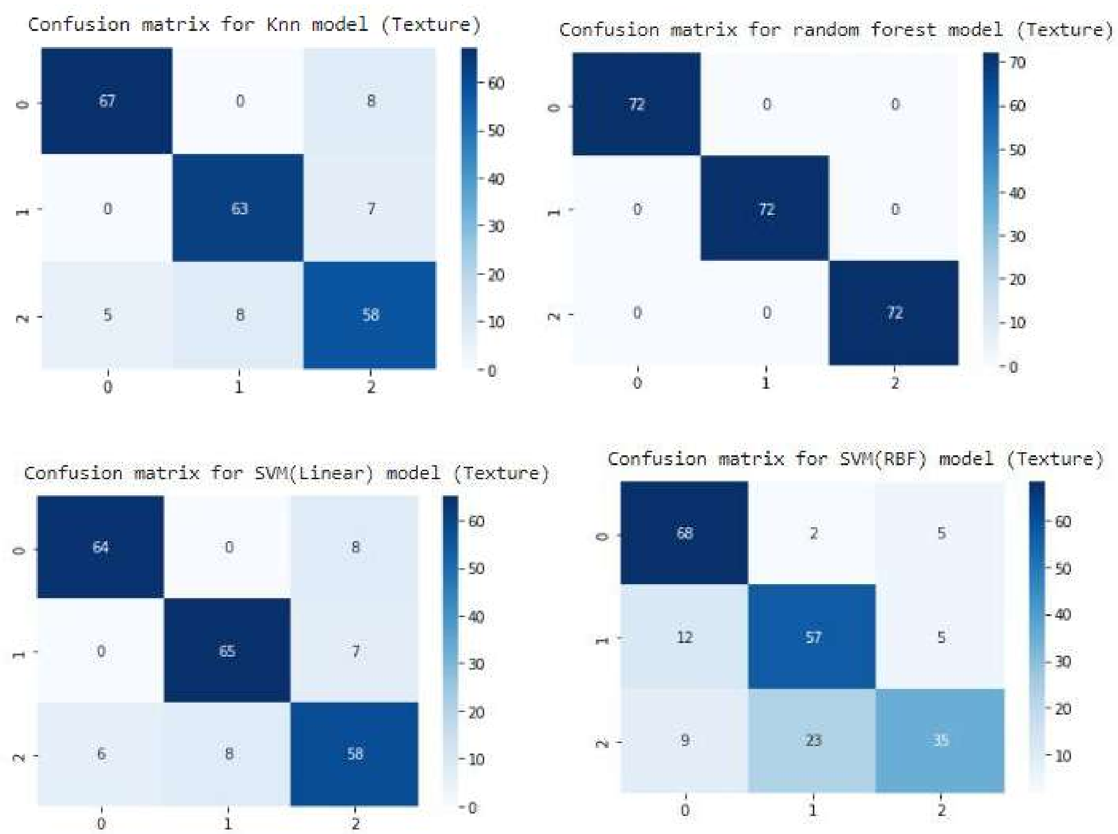


Figure 4.11 Confusion matrix (Training) of Texture feature

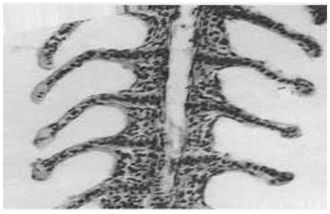
Table 4.10 Comparison between Performance measurements of Training (Texture)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	0.87	0.87	0.96	0.87	0.87
1	RFC	1.00	1.00	1.00	1.00	1.00
2	SVM(L)	0.87	0.87	0.96	0.87	0.87
3	SVM(RBF)	0.74	0.73	0.85	0.73	0.75

4.3.6 SHAPE (GLCM)

Level Cooccurrence Matrix (GLCM) method is a way of extracting second order statistical shape features. The approach has been used in a number of applications, Third and higher order textures consider the relationships among three or more pixels

(a) After



(b) Before

Contrast

$$\sum_{i,j=0}^{N-1} P_{i,j}(i - j)^2 \quad (19)$$

Correlation

$$\sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right] \quad (20)$$

Dissimilarity

$$\sum_{i,j=0}^{N-1} P_{i,j}|i - j| \quad (21)$$

Energy

$$\sum_{i,j=0}^{N-1} p_{i,j}^2 \quad (22)$$

Entropy

$$\sum_{i,j=0}^{N-1} P_{i,j}(-\ ln\ P_{i,j})\ (23)$$

Homogeneity

$$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}\ \ (24)$$

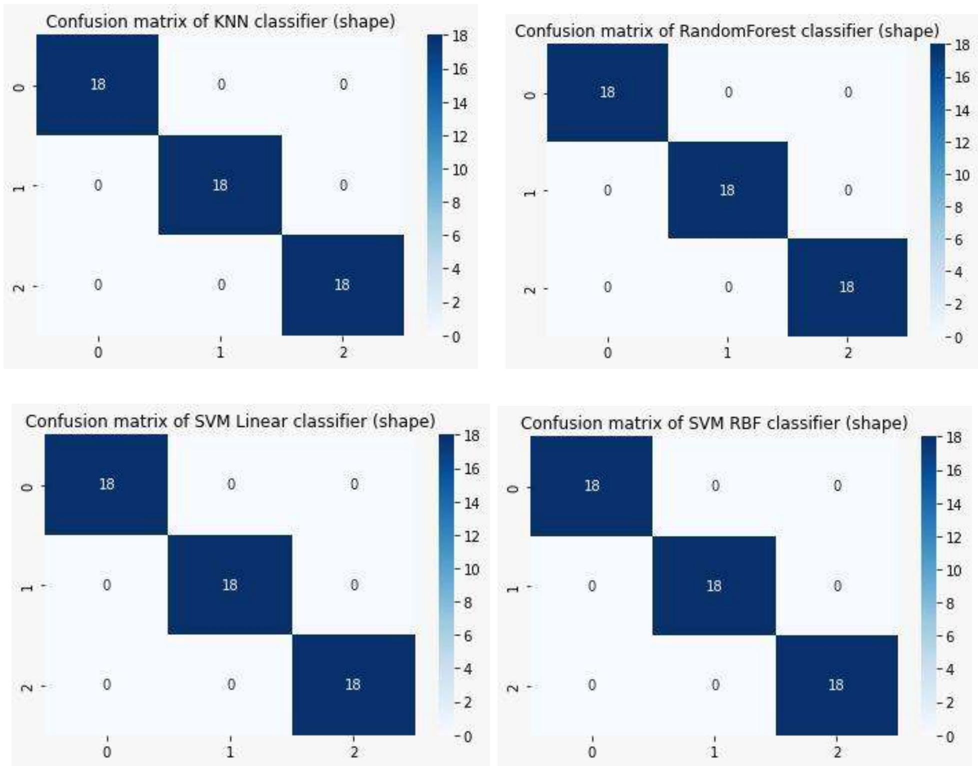


Figure 4.12 Confusion matrix (Testing) of GLCM feature

Table 4.11 Comparison between Performance measurements of training (GLCM)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	KNN	1.00	1.00	1.00	1.00	1.00
1	RFC	1.00	1.00	1.00	1.00	1.00
2	SVM(L)	1.00	1.00	1.00	1.00	1.00
3	SVM(RBF)	1.00	1.00	1.00	1.00	1.00

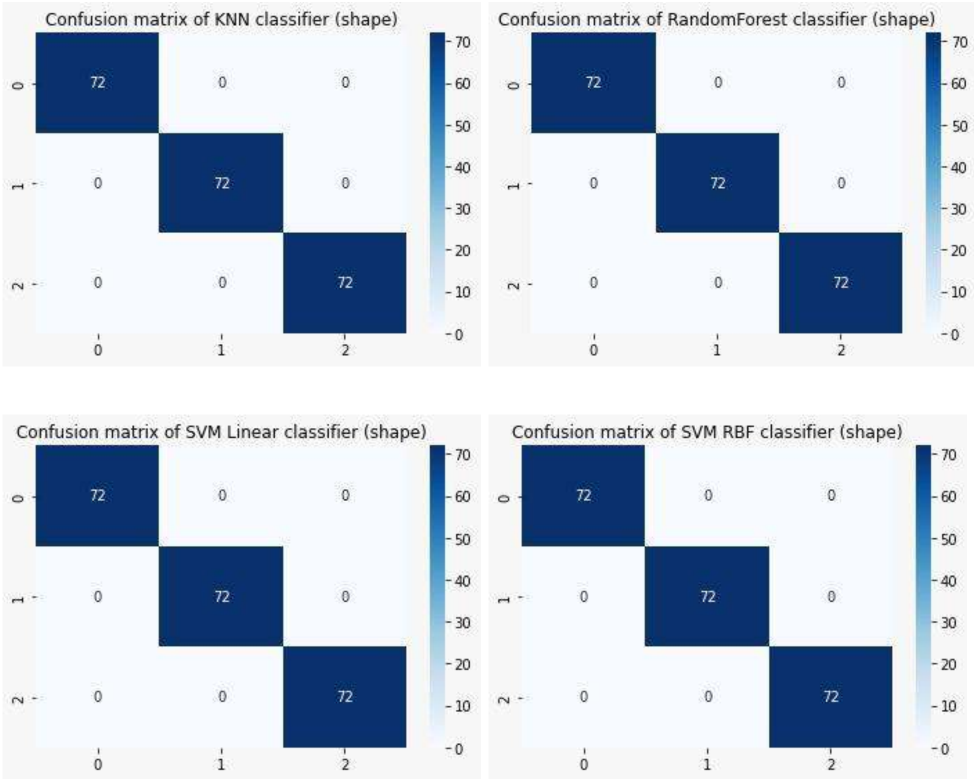


Figure 4.13 Confusion matrix (Training) of GLCM feature

Table 4.12 Comparison between Performance measurements of training (GLCM)

	classifier	Accuracy	Recall	specificity	f1-score	precision
0	**KNN**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
1	**RFC**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
2	**SVM(L)**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
3	**SVM(RBF)**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

4.4 Deep learning models

Deep learning uses artificial neural networks to perform complex calculations on large amounts of data. It is a type of machine learning that works on the basis of the structure and function of the human brain. But since the dataset is small, overfitting occurs when you achieve a good fit of your model on the training data while it does not generalise well on new, unseen data.

4.4.1 Convolutional neural network

The Convolutional neural network, CNN, or Convnet, is one of the most popular deep learning algorithms introduced in the 1980s by Yann LeCun , a postdoctoral computer science researcher. It is mainly used to mimic the human brain in the recognition and classification of various things, such as handwriting.The basic structure of CNN is given in Figure 4.14. [21]

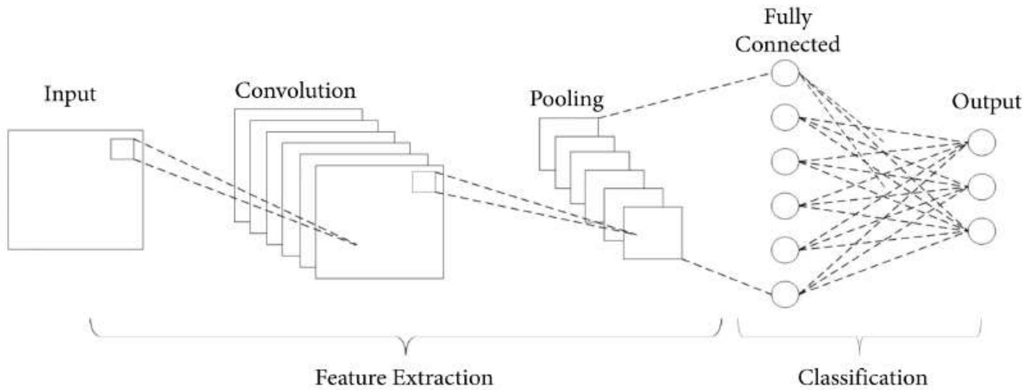


Figure 4.14 CNN structure [22]

We used 20 epochs for the CNN, the accuracy increased proportional to each epoch as shown in figure 4.15, we take the average of the accuracy history to calculate the training accuracy of the epochs as the following : **Training accuracy = 0.8096**
Also we calculated the loss of the training epochs it decreased proportional to each epoch as shown in figure 4.16, we take the average of the loss history to calculate the training loss of the epochs as the following: **Training loss = 0.243**
The CNN predicts all the testset correctly as shown in figure 4.17 which leads to overfitting due to our lack of data.

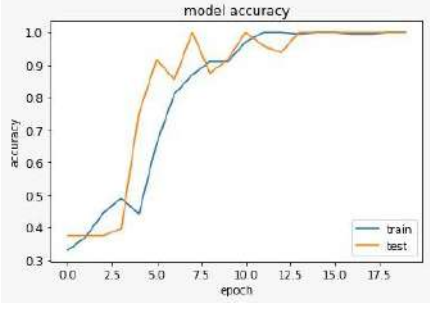


Figure 4.15. Plot of CNN training accuracy

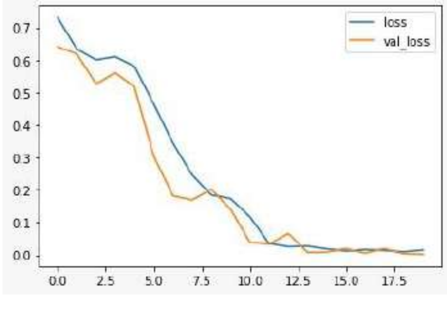


Figure 4.16. Plot of CNN training loss

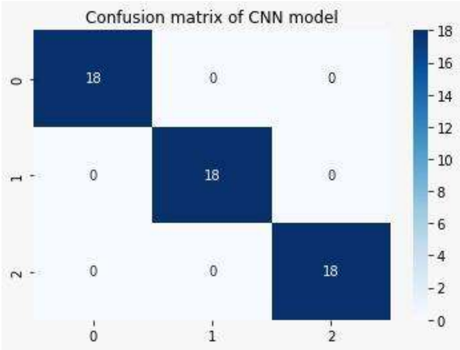


Figure 4.17 Confusion matrix of CNN model shows that the CNN model predicts all the given test set correctly so all the TP of the three classes are equal to 18 (The number of the test images in each class)

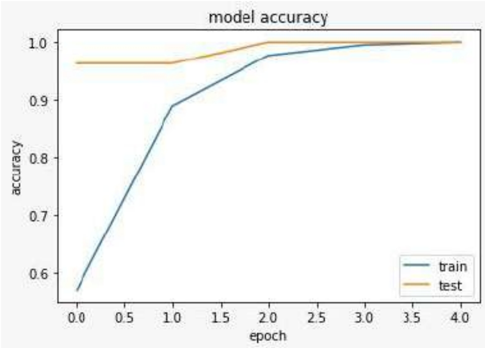
Table 4.13 Performance measurements of CNN model

Accuracy	Recall	specificity	f1-score	precision
1.00	1.00	1.00	1.00	1.00

4.4.2 Visual Geometry Group (VGG-16)

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition." The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to over 1000 classes[23].

We used 5 epochs for the CNN, the accuracy increased proportional to each epoch as shown in figure 4.16, we take the average of the accuracy history to calculate the training accuracy of the epochs as the following : **Training accuracy = 0.888**
Also we calculated the loss of the training epochs it decreased proportional to each epoch as shown in figure 4.17, we take the average of the loss history to calculate the training loss of the epochs as the following: **Training loss = 0.224 Training accuracy**
VGG-16 also predicts all the test set correctly as the CNN as shown in figure 4.18



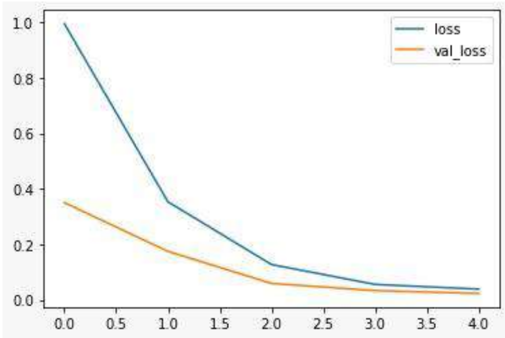


Figure 4.18. Plot ot VGG-16 training accuracy

Figure 4.19. Plot ot VGG-16 training loss

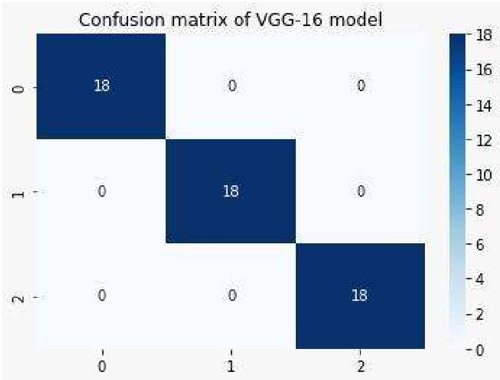


Figure 4.20. Confusion matrix of VGG-16 model shows that the CNN model predicts all the given test set correctly so all the TP of the three classes are equal to 18 (The number of the test images in each class)

Table 4.14 Performance measurements of VGG-16 model

Accuracy	Recall	specificity	f1-score	precision
1.00	1.00	1.00	1.00	1.00

Comparison between the algorithm of the testing set

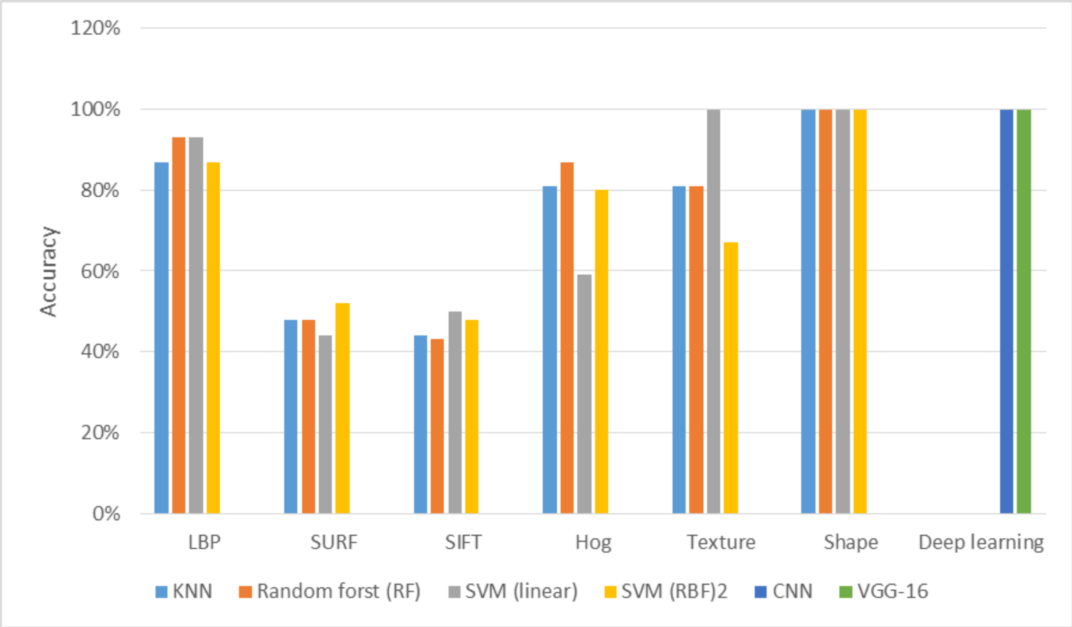


Fig 4.21. Comparison between the testing accuracy of all the used algorithm

Comparison between the algorithm of the training set

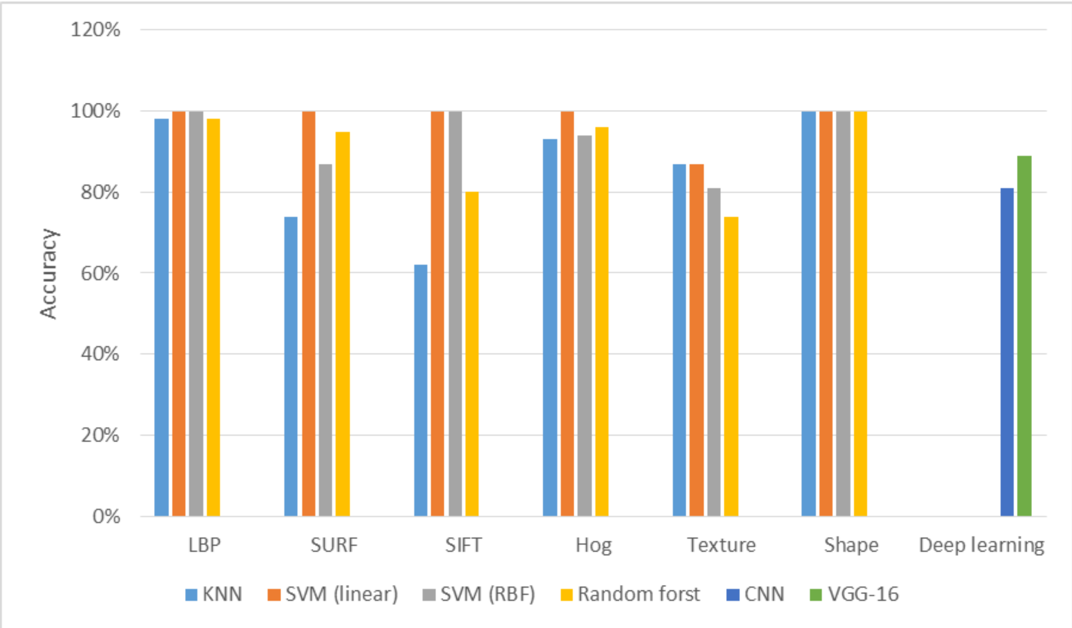


Fig 4.22. Comparison between the training accuracy of all the used algorithm

These comparisons show that the models are overfitting due to the lack of data.

Chapter Five

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Biomarkers are indeed an excellent tool to indicate a possible contamination in aquatic ecosystems. The results of determining Heavy metals concentrations in fish organs of Tilapia zilli indicated that Tilapia zilli gills contained the highest concentration (52.2%) of all the detected heavy metals.

In this study, the collection of fish gills samples from Abbassa farm, Abo-Hammad, Sharkia Governorate, Egypt has been done. The classification of 7 fish gills has been done with the help of experts. Manual labelling is time-consuming. However, the automation of classification saves time and, at the same time, can be very accurate in the prediction of water quality as an expert doing manual identification may do something wrong and thus affect the results. By applying augmentation, we obtained 270 samples just from 7 sample images by attaining 90 images out of each class. According to the results of augmentation obtained, we applied feature extraction by using 6 features which are (colour (HOG), shape (GLCM), texture (Gabor filter), LBP, SURF, SIFT), it turned out that both LBP and shape features are the most effective features for classification accuracy. By applying the 4 classifiers which are (KNN, Random Forest, SVM Linear, SVM RBF) on these 6 features, it was found that the random forest (RF) is the most suited algorithm for classifying the water quality according to the fish gills. In deep learning, it was discovered that both the convolutional neural network (CNN) and Visual Geometry Group (VGG-16) have the highest accuracy as they are overfitting because of a lack of data. Deep learning can be more useful when dealing with large amounts of data.

The results of water quality may be unrealistic, so we have to train the model more than once to make the program results more realistic.

This research is intended to highlight the capabilities of usage of image classification techniques based on fish gills as biomarkers for water quality. We aim to improve the understanding of fish gills as biomarkers and effects of fish on water quality.

5.2 Achievements

Throughout the year, we’ve submitted to multiple competitions, we got accepted in some and we couldn’t make it in the others.

First, we submitted in the Dell Technology competition; we made it through but we weren't accepted in the design stage.

We submitted in the Academy of Scientific Research and in the ITAC Program, but we weren't accepted. We also submitted in the AAST rally, we got accepted in the early stage but we couldn't make it.

In UAE we successfully passed the filtration stage, but weren't accepted past this stage. Finally, we submitted it to AWS. We were among 40 teams out of 400 that got accepted. Besides we were the only team representing AAST. We made it to the final stage but we didn't win, though we were in fourth place in our track.

5.3 Future work and suggestion

The weaknesses and limitations of each of the tools and techniques developed in the research study have indicated the following areas as recommendations for further work:

One of the best steps that we can take is to turn this idea into a project, as water is an essential source of everyone's life and the level of health of the organism is linked to the percentage of water purity and its suitability for drinking.

Providing this project on a large scale will give organisations interested in studying water quality a great opportunity to facilitate this process.

We are working to expand our project and publish it to all specialised agencies.

There will be two options. We will provide a complete package that contains the software and a microscope compatible with the software as the first option. The second option will be the software only. It will be an affordable price, especially if a person has a compatible microscope or knows somewhere that has this smart microscope without forcing them to buy a new microscope.

To implement the project functions in a more professional way, we will make a camera software that will be a replacement for the microscope, which will cost more time to develop, as shown in Figure 5.1.

This helps in speeding up production to cooperate with the competent bodies that can help provide test samples and are interested in research in this field. We hope to present the project to the Water Research Foundation.

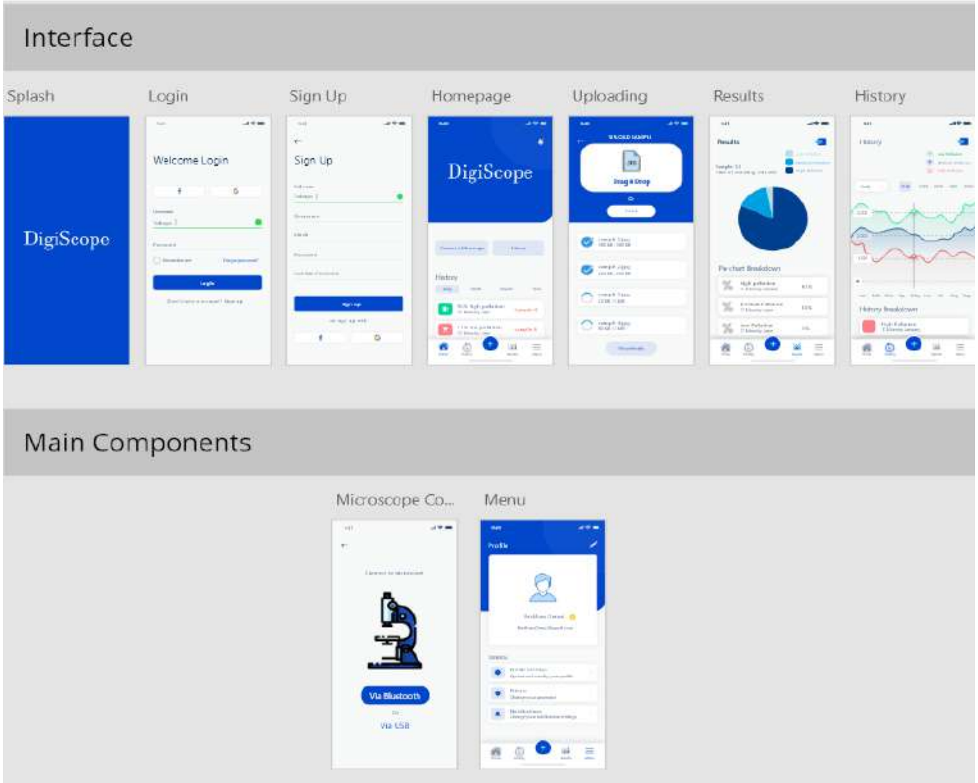


Figure 5.1 Initial prototype

REFERENCES

[1] Marabelle Delaurentis, MARCH 25, 2021, CIVIC ISSUES BLOG, MAY 23, 2022, <https://sites.psu.edu/marabelleolivia/2021/03/25/why-should-we-turn-off-the-tap-water-when-brushing-our-teeth/>

[2] Haresh Bambhaniya, *Engineering*, engineeringchoice.com/water-pollution/#:~:text=is%20water%20pollution%3F-,Water%20pollution%20is%20the%20contamination%20of%20water%20sources%20by%20substances,their%20way%20to%20the%20water.

[3] Erin Spencer, JANUARY 17, 2020, *Ocean Conservancy*, march 2020, <oceanconservancy.org/blog/2020/01/17/gills/#:~:text=Gills%20are%20branching%20organs%20located,that's%20dissolved%20in%20the%20water>

[4] Thiago E.M. Parente and Rachel Ann Hauser-Davis,December 2013 ,“Pollution and Fish Health in Tropical Ecosystems”,*ResearchGate* ,March 2014,Chapter (7), p.p. 164-181 ,researchgate.net/publication/258113051_The_Use_of_Fish_Biomarkers_in_the_Evaluation_of_Water_Pollution_In_Pollution_and_Fish_Health_in_Tropical_Ecosystems_ISBN_9781482212877_CRC_Press.

[5] Eneji, I.S.; Ato, R.SAnnune, P.A.*IAEA*,2011,2011,inis.iaea.org/search/search.aspx?orig_q=RN:43028008

[6] H. Fernandez-Canque ,B.J. Beggs,January 2006,ResearchGate,.researchgate.net/publication/242298852_MicroOrganisms_Detection_in_Drinking_Water_Using_Image_Processing

[7] Anaahat Dhindsa, Sanjay Bhatia , Sunil Agrawal and B.S. Sohi, September 2020,*ResearchGate*,January 2021,researchgate.net/publication/346369631_An_Efficient_Microbes_Detection_System_using_Microscopic_Images_via_Morphological_and_Correlation_Based_Features

[8] Yue Liao,Jian-yu Xu,Zhu-wei Wang,April 2012,*ResearchGate* july 2012,researchgate.net/publication/223980084_Application_of_biomonitoring_and_support_vector_machine_in_water_quality_assessment

[9] Shivam Sharma ,May 15, 2021,Analytics vidhya,analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/

[10] JavaTPoint,javatpoint.com/machine-learning-support-vector-machine-algorithm

[11] Sruthi E R,June 17, 2021,*Analytics Vidhya*,July 2021,analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[12] dataschool contributors,March 25, 2014,*Data School*,june 2014,dataschool.io/simple-guide-to-confusion-matrix-terminology/

[13] Fatih Demir, in Artificial Intelligence-Based Brain-Computer Interface, 2022,*ScienceDirect*,sciencedirect.com/topics/engineering/confusion-matrix

[14] Şaban Öztürka *, Bayram Akdemirb,2018, ScienceDirect,sciencedirect.com/science/article/pii/S1877050918307890

[15] Mrinal Tyagi,Jul 4, 2021,*Towards Data Science*,Augast 2021,towardsdatascience.com/hog-histogram-of-oriented-gradients-67ecd887675f

[16] Michael Hahnle,Frerk Saxen,Matthias Hisung,K. Doll,June 2013,*ResearchGate*,october 2013,researchgate.net/figure/Calculation-steps-of-the-HOG-descriptor_fig1_261285146

[17] Herbert Bay,Tinne Tuytelaars,Luc Van Gool,July 2006,*ResearchGate,December 2006,https://www.researchgate.net/publication/225761164_SURF_Speeded_up_robust_features*

[18] Leqi Jiang, ... Xingming Sun, in Advances in Computers, 2021,*ScienceDirect,*sciencedirect.com/topics/computer-science/scale-invariant-feature-transform

[19] A. Humeau-Heurtier, "Texture Feature Extraction Methods: A Survey," in *IEEE Access*,6 June 2022 ,vol. 7, pp. 8975-9000, 2019, doi: 10.1109/ACCESS.2018.2890743.

[20] Aishwarya Singh,August 29, 2019,*analyticsvidhya,*6 June 2022,analyticsvidhya.com/blog/2019/08/3-techniques-extract-features-from-image-data-machine-learning-python/

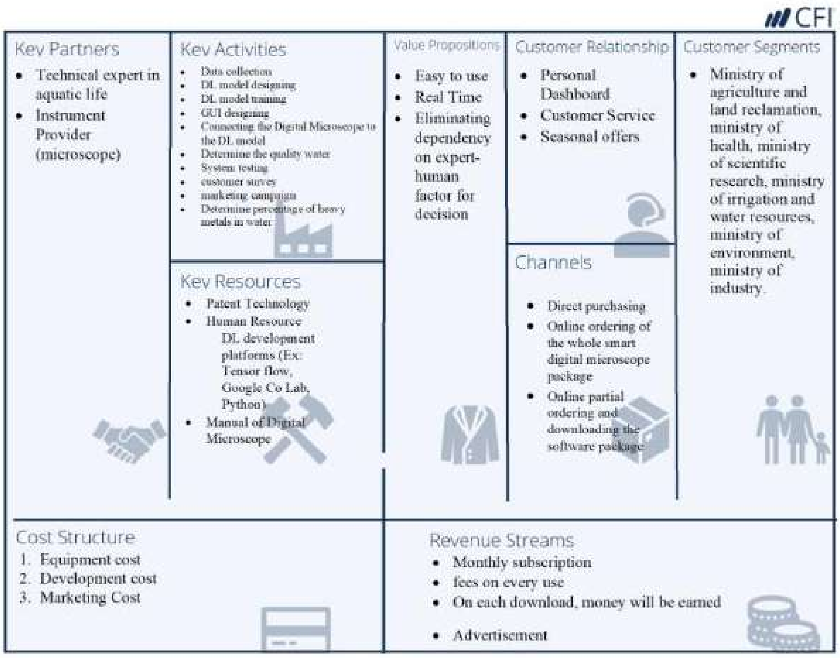
[21] Ben Dickson,January 6, 2021,TechTalks,March 2021,bdtechtalks.com/2020/01/06/convolutional-neural-networks-cnn-convnets/

[22] MK Gurucharan ,DEC 7, 2020,*UpGrad*,March 2021,upgrad.com/blog/basic-cnn-architecture/

[23] Popular networks,20 November 2018,Neurohive,January,2018,

/neurohive.io/en/popular-networks/vgg16/

APPENDICES



Appendix A Business Model Canvas