# Cancer Type Identification Through Parameterization of Image Patterns Using Machine Learning

**By**

**SARAH TARIQ SHEIKH**

**Reg # 02282111003**

**National Center for Bioinformatics**

**Faculty of Biological Sciences**

**Quaid-I-Azam University,**

**Islamabad, Pakistan**

**2023**

# Cancer Type Identification Through Parameterization of Image Patterns Using Machine Learning

**By**

**SARAH TARIQ SHEIKH**

**Reg # 02282111003**

A thesis submitted in the fulfillment of the requirements for the degree

of

MASTER OF PHILOSOPHY

In

BIOINFORMATICS

**National Center for Bioinformatics**

**Faculty of Biological Sciences**

**Quaid-I-Azam University,**

**Islamabad, Pakistan**

**2023**

# CERTIFICATE

This Thesis is submitted by **Miss Sarah Tariq Sheikh** from National Centre for Bioinformatics, Faculty of Biological Sciences, Quaid-i-Azam University Islamabad, Pakistan, is accepted in its present form as satisfying the thesis requirement for the Degree of Master of Philosophy in Bioinformatics.

Internal Examiner: ______________

Dr. Sajid Rashid

Professor & Supervisor

Quaid-i-Azam University Islamabad

External Examiner: ______________

_______________

_______________

_______________

Chairman: ______________

Dr. Syed Sikandar Azam

Professor

Quaid-i-Azam University Islamabad

Date: 16th February, 2023

# DECLARATION

The work reported in this study, entitled **"Cancer Type Identification Through Parameterization Of Image Patterns Using Machine Learning"** was carried out by Sarah Tariq Sheikh under the supervision of Dr. Sajid Rashid, from National Centre for Bioinformatics, Faculty of Biological Sciences, Quaid-I-Azam University, Islamabad, Pakistan. Hereby, I declare that the title of the thesis and all the contents presented in the following study are products of my own effort and no part has been copied from any published source (except the references, standard mathematical or genetic models /equations /formulas /protocols,etc.). None of the work has been submitted for award of any other degree /diploma. The University may take action if the information provided is found inaccurate at any stage.

**SARAH TARIQ SHEIKH**

# PLAGIARISM CERTIFICATE

It is certified that **Sarah Tariq Sheikh** (02282111003) has submitted her M.Phil. Dissertation entitled **“Cancer Type Identification Through Parameterization Of Image Patterns Using Machine Learning”** that has been checked on Turnitin for similarity index (plagiarism).

Overall plagiarism = 14% that lies in the limit provided by HEC (19%)

**Dr. Sajid Rashid**

**Professor**

**National Center of Bioinformatics**

**Quaid-i-Azam University, Islamabad**

# DEDICATION

I dedicate this dissertation to my beloved parents and husband.

# ACKNOWLEDGEMENTS

I would like to start by thanking my supervisor **Dr. Sajid Rashid**, for his moral support, guidance, enthusiastic encouragement and valuable critique for this research work. I am forever inspired by his ideas, devotion, and dedication. I extend my appreciation to **Dr. Amir Ali Abbasi** and **Dr. Syed Sikander Azam**, Chairman of the National Centre for Bioinformatics at Quaid-I-Azam University, Islamabad.

I am grateful to **Amana Batool**, **Hafsa Khalil** and **Muhammad Shafique**, my mentor and lab colleague for support and aid in tackling problems and errors. Bundle of thanks to my amazing and supporting lab colleagues **Zainab Kousar**, **Maha Yousaf** and **Iqra Kamal** and all my lab fellows for providing a stimulating and constructive environment in the lab.

I am forever indebted to my parents for providing me with the means to be able to get to this level. I owe this degree to my **Parents**; they sacrificed their own wishes to fulfill my dreams. I am thankful to my brothers (**Ahmed Tariq Sheikh**, **Asjad Tariq** and **Moeen Tariq**) and sister (**Amna Tariq Sheikh**) for guidance and encouragement. I am also thankful to my sister in laws (**Sana Khalid**, **Huda Farooq** and **Uma Kalsoom late**) for unconditional support and wishes.

I would like to thank my husband **Ahmed Belal Hashmi** for providing me with an environment at home where I was able to complete this project. You really helped me out throughout this project through guidance, encouragement and love. I am forever indebted to you. I am also thankful to my father in law **Belal Muhammad Hashmi** for his ultimate support and guidance.

I would like to thank my best friends **Zainab Kousar**, **Atika Khan** and **Laila Sehar** without whom this journey would have been difficult and boring. Thank you for the ultimate support, laughter, adventure, memories and to prove that friends can be a family too.

And also, the First and the Last, the Kind and the Just, to Him I owe everything.

**Sarah Tariq Sheikh**

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| K-NN | K-NEAREST NEIGHBOR |
| ML | Machine Learning |
| DNA | Deoxyribonucleic acid |
| CLAHE | Contrast Limited Adaptive Histogram Equalization |
| ROI | Region of interest |
| CT | Computed Tomography |
| MRI | Magnetic Resonance Imaging |
| ROC | Receiver Operator Characteristic |
| AUC | Area Under The Curve |
| Stdev | Standard Deviation |
| CAD | Computer Aided Diagnosis |
| JPEG | Joint Photographic Experts Group |

# ABSTRACT

Finding patterns in the histopathology images of cancer is the new challenge of this era. The traditional method for scanning histopathology images is done manually by pathologists who look for patterns in whole slides. This method is quite time taking and laborious as it takes days and even months for proper diagnosis. Moreover, there is a chance of human error in this approach. Now, many machine learning algorithms like K-NN, logistic regression, random forest, decision trees etc. are currently being used in healthcare and image processing to help doctors and scientists to diagnose diseases faster. The goal of this study is to develop improved strategies for various CAD phases that will play a critical role in not only minimizing the variability gap between and among observers, but also reduce the overall time and cost. The dataset of colon cancer was taken from the Kaggle database. The histopathology images were 768x768 pixels in size and in JPEG format. The images were first observed manually to find any differences between benign and malignant images. The nucleus regions were quite different between the two, so the nucleus regions were selected as the regions of interest in this study. Different preprocessing steps, for example, brightness and contrast normalization, were performed to increase the quality of images for analysis. Through color segmentation only nucleus regions were extracted, while masking the other features in the images as white. Many features of the nucleus were extracted, like nucleus mean area, nucleus area standard deviation, nucleus mean height, nucleus mean width and aspect ratio and were output to a data file. This file was further used in the training and testing phase using the K-NN model due to its simplicity. Through graphical and statistical analysis, it was observed that the nucleus mean area, mean height, aspect ratio and nucleus area standard deviation were quite higher in malignant images than benign images. Through correlation between the features and malignancy, it was further reinforced that the nucleus area mean, nucleus aspect ratio and nucleus height had more effect on malignancy than other features. Hence, during the training phase only these features were selected, while dropping the others from the feature vector. A total of 1090 images were used in this study, split 80/20 for training and test phases respectively, resulting in 870 images for training and 220 for the testing phase. Different values of k were selected to find the best value where accuracy of the model exceeds. The model gave 90.91% accuracy at k=8,

which yields an area under the ROC curve of 0.87, which indicates a good performance of the model. This software can be used for early screening purposes due to its high accuracy, which will not only help in diagnosis at a faster rate but also can be made as a standard technique used in cancer diagnosis.

# INTRODUCTION

# 1. INTRODUCTION

Cancer has been one of the biggest challenges for medical science to solve in the past century or so. We do not yet possess reliable tools to battle cancer including diagnostics and treatment. A significant effort from the scientific and medical community has been made to change that in modern times but human beings have yet to solve this long-standing issue. Every year there are incalculable resources spent into finding the true cause, management regime and cure to this disease but it is still a work in progress. According to the World Health Organisation (WHO) the family of cancer diseases is a leading cause of death worldwide accounting for nearly ten million deaths in 2020 or nearly one in six deaths (Ferlay et al., 2021). With the latest advancements and improvements in technology humans now have better tools at our disposal to aid in this venture. This research is focused on the use of artificial intelligence and machine learning, a major breakthrough in technology, as the basis of diagnosing cancer and making a machine-assisted prognosis/diagnosis for any prospective patient.

An average machine learning model is enough for surveying and classification of biological data. Recently, a large quantity of image data has been produced by neuroscientist Steve Finkbeiner through robotic microscopy to study brain cells (Webb, 2018). But the computational cost was too high to properly analyze this data. Finkbeiner welcomed the idea to use Google technologies for accelerating scientific discoveries. Short of using cloud-based machine learning infrastructure this research attempts to recreate the same capabilities albeit in a local environment and at a smaller scale.

## 1.1 Machine learning

The use and development of computer systems that can learn and adapt without being given specific instructions through analysis of data patterns and making insights from them using algorithms and statistical models is termed machine learning (Mahesh, 2019). ML has a number of uses in a variety of industries, from healthcare to natural language processing. This revolution affects all fields, including biology and bioinformatics. Prior to the development of machine learning, these fields were unable to derive any usable information from large biological data sets. However, as of now, ML techniques, like deep learning, can discover properties of complicated datasets and present them in an comprehensible manner. Even though this technology has been around for a while, it has only recently been applied to the analysis of enormous amounts of data.

Since image processing is such a helpful technology, there is an increasing need for it in disease diagnosis each year. Machine learning-based image processing was first developed in the 1960s in an effort to stimulate the human visual system and automated image analysis (Huang et al., 2022). Solutions for certain problems started to emerge as technology advanced. Trying to give computers intelligence seems to be both a challenging and impossible task, yet it is actually reasonably easy. It can be broken down into the following seven essential steps given below.

### 1.1.1 Collecting Data

Machines initially learn from the data that is given to them. It is essential to collect reliable data in order for the machine learning model to discover patterns with accuracy. The quality of data used to feed the model will determine how accurate it is. The model will produce inaccurate results or unreliable forecasts if the data is wrong or out-of-date.

### 1.1.2 Preparing the Data

When necessary information has been acquired, the preparation phase begins. This is done by combining and randomizing the order of data in the dataset. This method ensures it is made sure that the distribution of the data is matched and that the ordering won't impact how the data are trained from or incorporate bias into the model. The process of data cleaning, which is the following phase, entails getting rid of unnecessary details, incomplete data, redundant values, rows, and columns, as well as altering data types. The dataset's rows, columns, or index of rows and columns may occasionally need to be changed to fit the model's specifications for format. For understanding the structure of the data and the relationships between the multiple variables and classes it contains, visualization of the data helps to do so. The cleaned data is divided into a training set and a test set. The training set is the set from which the model learns. The test set is used to verify the effectiveness and accuracy of the model.

### 1.1.3 Choosing a Model

After any technique has been applied to the gathered data, a machine learning model determines the result. It's crucial to choose a model that works best for the current issue. For a variety of tasks, including speech recognition, image recognition, event prediction, etc., scientists and engineers have developed a number of models over the years. Choosing whether the model performs best with qualitative or quantitative data is also important.

### 1.1.4 Training the Model

The most crucial phase of ML is training. During training, the training data is fed into the machine learning model to search for patterns and obtain conclusions. As a consequence, the model gains knowledge from the data about which patterns lead to which results. With further training the model gets more accurate at anticipating outcomes over time.

### 1.1.5 Evaluating the Model

Once the training is done, it is essential to check if the model is operating correctly. To do this, effectiveness of the model is assessed using novel data. The unobserved data is taken from the previously separated testing set. Using the same data for testing would not produce accurate results because the model is already familiar with the data and recognises the same patterns in it as it did during training. This will produce excessively high accuracy. This will give a more accurate evaluation of the model's performance and speed when applied to test data. Once the test data is provided to the ML model the predicted outcomes are compared with already known outcomes for the test data (which may have been calculated manually ahead of time) to arrive at the accuracy score of the trained model.

### 1.1.5 Parameter Tuning

Once the model has been created and inspected, the next task is to determine if there is any way to improve its accuracy. To accomplish this, the model's parameters are tweaked. The variables that are selected by the programmer are the model's parameters. When a certain parameter value is reached, accuracy will reach its highest. Parameter tweaking (or tuning) is the method used to find these settings.

### 1.1.6 Making Predictions

Last but not least, the chosen model may be applied to future data to generate accurate forecasts.

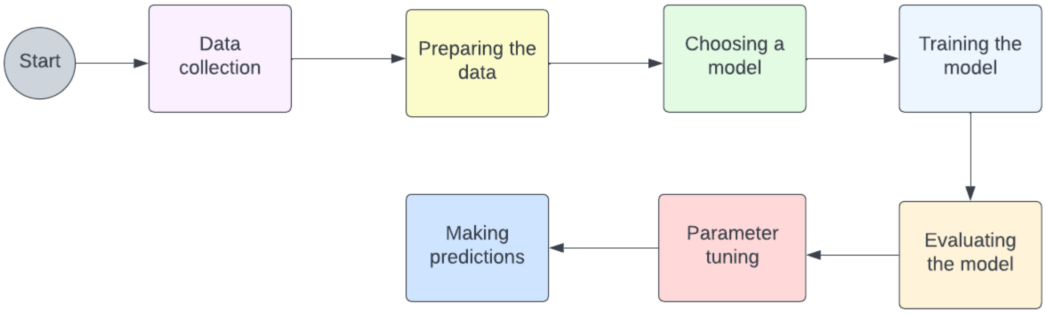


**Figure 1.1 Machine learning process.**

## 1.2 Image recognition technology

The concept of "image recognition" refers to how people analyze and understand visual information in the form of images. Many medical disorders are subcutaneous (not superficial) and cannot be visually inspected without assistance so different tests like computed tomography (CT) scans, magnetic resonance imaging (MRI) and others are used to augment the diagnostic toolkit. As the size of medical image dataset output from these methods is growing it becomes increasingly challenging for the medical institutions and practitioners to analyze such vast amounts of data manually (Zhang & Xin, 2016). However, many medical professionals are currently limited to analyzing these medical images alone. It is simple to misinterpret medical imaging due to medical staff having an excessive amount of subjective judgements about medical quality, knowledge level, and personal competence (Long et al., 2016). Machine learning-based image processing technology offers a technique to efficiently reduce the impact of human error. Image recognition is a critical stage in the computer's processing of visual data (Motofumi et al., 2016). Once successfully detected, the visual information can be analyzed and interpreted (Komura & Ishikawa, 2018). Currently, machine learning technology for image processing has an accuracy rate of more than 99%. Even though the accuracy rate is very high, it's still crucial to have the results verified by a qualified professional.

Four key steps in image recognition are:

- Image information acquisition
- Image information processing
- Feature extraction
- Category creation
- Classification

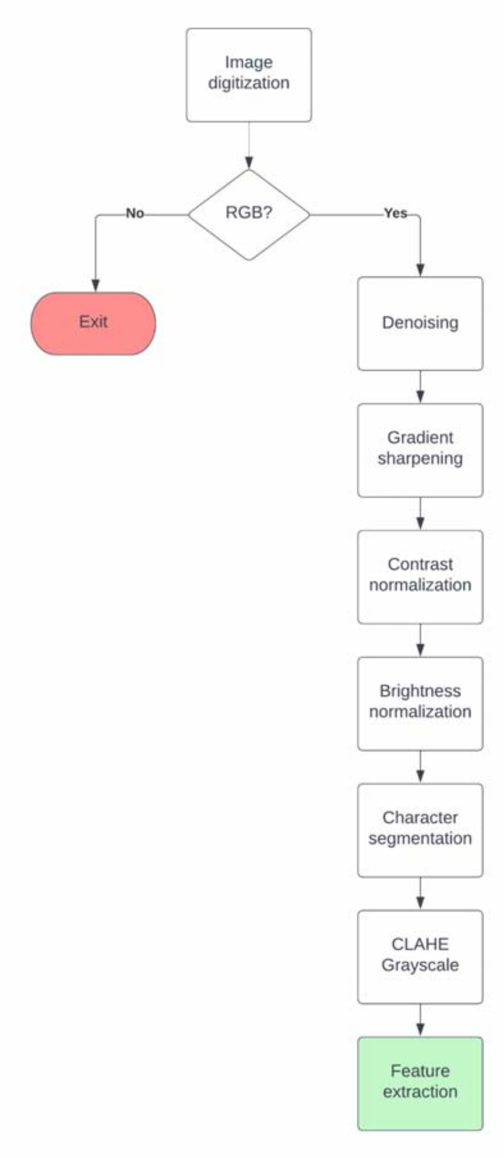


**Figure 1.2: Picture information preparation.**

### 1.2.1 Digital images

The picture elements that make up digital photos are called pixels. Each pixel represents the smallest dot of a given color. These dots are arranged together in order to form the complete photo. Ordinarily, pixels are organized in an orderly, rectangular array.

### 1.2.2 Size of image

The size of an image is determined by dimensions of this pixel array. The width and height of the image correspond to the column and row counts of the array. As a result, the pixel array is a matrix with M columns and N rows.

### 1.2.3 Pixel

Within the picture matrix, a pixel is identified by its coordinates at x and y. X and y axes of the image matrix coordinate system are said to increase from left to right and from top to bottom, respectively.

### 1.2.4 Brightness and Colour

Each pixel has a distinct intensity value, also called brightness. If every pixel has the same value, the image will always be a uniform shade of black, white, gray, or another color. The intensity type used for each pixel determines how distinct image types differ from one another. Only the deepest (black) and lightest (white) levels of intensity are available in black-and-white images. On the other hand, the intensity of color images is determined by the darkest and lightest tones of the three colors: red, green, and blue (Figure 1.3). The various combinations of these color intensities result in a coloured image.The two most basic categories of digital images, which are black and white and coloured, are known as grayscale and RGB (red, green, blue) images (Ketcham et al., 1974).

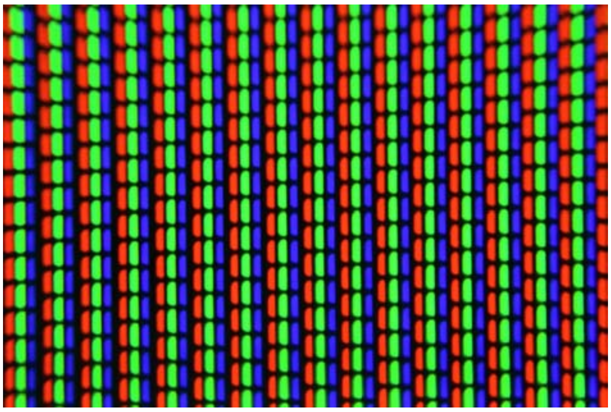


**Figure 1.3: Physical pixels on a TV screen.**

### 1.2.5 Data Storage

Bits are used to define the intensity levels in digital photographs. The only two potential values for a binary bit are 0 or 1. There are 256 potential values in the 8-bit intensity range (0 to 255). Standard digital photos have an 8-bit range of values, while B&W images have a single 8-bit intensity range and RGB images employ 8-bit intensity ranges for each color. RGB images are also known as 24-bit color images since they have three 8-bit intensities. In other words a single pixel in a grayscale image would be stored in 8 bits (or 1 byte) whereas a single pixel in a coloured image (without transparency) would require 24 bits (or 3 bytes).

### 1.2.6 Resolution

An extra component required for a digital image to accurately depict the real world is resolution. Resolution is the measure of how closely spaced apart an image's pixels are and how small of an area a single pixel represents. If a picture has two versions, one with high resolution and the other with low resolution, the high resolution picture will have more pixels and display more detail. The picture with low resolution will have an overall lower count of pixels and subsequently a lower level of visual detail (Figure 1.4).

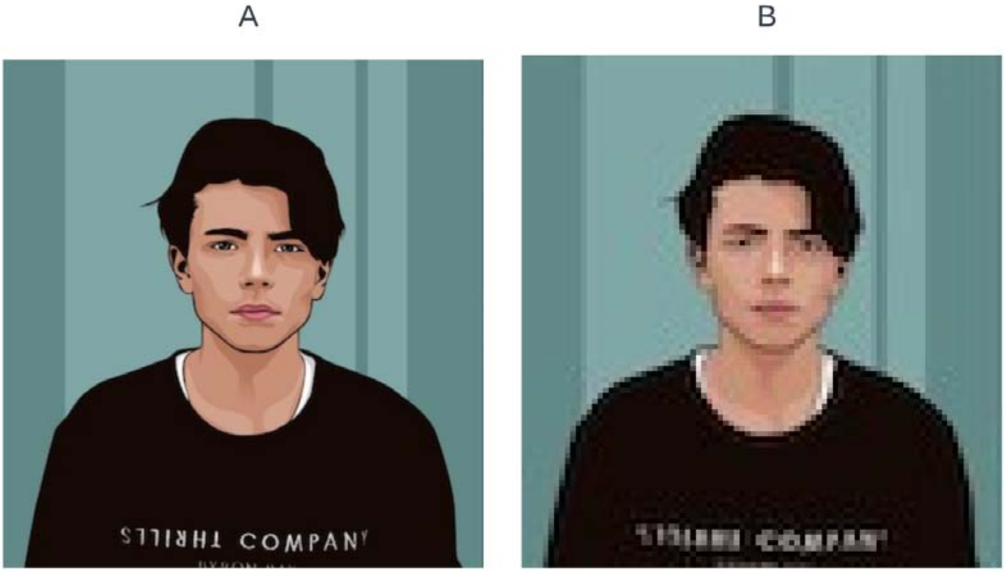


**Figure 1.4: High (A) and low (B) resolution image comparison.**

### 1.3 Digital Image Processing

With a computer, digital image processing entails editing digital photos. It is a subset of signals and networks that places a high level of importance on images and visual data. The creation of a computer system that can process images is the main objective of digital image processing. The system accepts a digital image as input and processes it using efficient algorithms to create a final image (Ketcham et al., 1974). Adobe Photoshop is one of the most widely used software programmes for manipulating digital images.

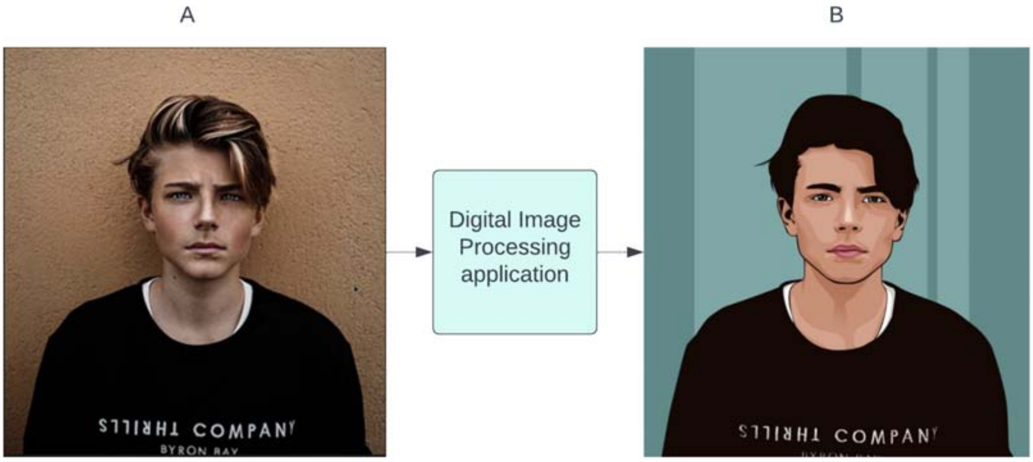


**Figure 1.5: Digital image processing posterization effect.**

In the above figure (Figure 1.5), a digital image (A) has been sent to a digital system to convert a man’s photograph to look like a poster (B).

- Basically, image processing involves the following three steps:
- Using picture acquisition techniques to import the image
- Examining and modifying the picture
- Output from image analysis that has the potential for changing the final image or statement

## 1.4 Histopathology based stained images

Early cancer diagnosis and treatment can lower mortality rates and raise long-term life expectancies. Routine screening, laboratory testing, imaging tests (CT, MRI, PET, X-ray), and biopsy are all used in the traditional diagnosis of cancer (L et al., 2012) (Mahmood et al., 2020).The most effective diagnostic method for obtaining tissue samples from an afflicted area is a biopsy. Histopathology is the study of tissue sections under a microscope for medical purposes after they have been stained. Pathologists frequently utilize hematoxylin and eosin (H&E) staining to distinguish tissues. Hematoxylin staining causes the cell nuclei to appear purplish blue, and eosin staining causes the cytoplasm and connective tissue to appear pink (L et al., 2012). Analyzing histopathological images is done to look at cell shape and tissue growth patterns. Pathologists use various magnifications of a microscope to examine tissue slides. At modest magnifications (4X, 10X), they can see the entire tissue sample and gain insight into the malignant region, particularly the malignancy's architectural layout. Then, to assess the morphology at the cellular level, they carefully examine the slides at greater magnifications (40X, 60X, 100X, 200X, and 400X) (He et al., 2012). Each malignant tumor typically exhibits certain microscopic characteristics, such as aberrant cell shape and size, hyperchromatic (dark) nuclei, conspicuous nucleoli, increased nucleus to cytoplasm ratio (NC ratio), increased number of mitotic figures, and abnormal mitotic divisions(L et al., 2012). Large cells, vesicular nuclei, and an abundance of eosinophilic cytoplasm are a few of the microscopic characteristics (L et al., 2012) (Rivera & Venegas, 2014). Early stages of SCC can be seen to include intercellular bridges, which are a feature of squamous epithelium (Nguyen et al., 2015). Cell level data allows for the differentiation of histopathological variations of SCC and ADC (Zhang et al., 2015).

Manual histological examination requires a lot of effort, takes a long time, and is subjective, with expert variances both across and among observers(L et al., 2012) (Mahmood et al., 2020). Clinical decision-making needs to be automated as a result of these constraints. In order to generate speedy and reproducible results, some researchers have created computer-aided diagnostic (CAD) systems for carcinoma diagnosis and categorization utilizing a combination of image processing, pattern recognition, machine learning (ML), and deep learning (DL) techniques (Komura & Ishikawa, 2018).

## 1.5 Image preprocessing

Before developing the model, most computer systems analysts spend a significant amount of time performing data pretreatment, also known as data cleaning. Outlier detection, missing value treatments, and deleting undesired or noisy data are a few examples of data preparation. Similar to that, "image preprocessing" refers to actions taken on photographs at their most basic level of abstraction. If entropy is a measure of information, then these actions diminish rather than increase the information content of the image. Preprocessing aims to improve the image data by reducing unwanted distortions or improving certain image properties important for further refining and examining tasks (Kumar et al., 2020).

### 1.5.1 Image size normalization

The process of normalizing involves scaling up or down photos of various sizes. A minimum bounding box is applied to the character, and the element is cropped and then enlarged to fit into a window in order to bring all characters into a standard size platform in order to extract features on an equal footing. The cropped element is normalized without changing the aspect ratio to the appropriate size, which is typically 36x36, 100x100, 512x512, 768x768 or 1024x1024 pixels (Taylor et al., 2013).

### 1.5.2 Image magnification

A pathologist often examines tissue samples under a microscope using a variety of magnifications to analyze and make a diagnosis. Compared to high magnification photos, low magnification photographs have a wider field of vision. Researchers extract features using various magnifications for their experiments. For instance, Hosseini et al. developed a fractal-based strategy to discern between normal and esophageal dysplasia by training a classification model separately using pictures from 10X and 20X magnifications. The role of magnification in categorization of cancer was investigated by Kumar et al. At 4X and 10X magnification, they were more accurate, but at 40X, they were less accurate.

### 1.5.3 Image enhancements

Digital images are adjusted throughout the process of image enhancement to provide outcomes that are better suited for display or additional image analysis (Hummel, 1977). For instance, eliminating noise, sharpening, or brightening an image to make it simpler to spot important details (Kumar et al., 2020).

### 1.5.4 CLAHE (Contrast Limited Adaptive Histogram Equalization)

Contrast over-amplification is addressed by CLAHE, an adaptation of Adaptive Histogram Equalization (AHE). CLAHE operates on specific parts of the picture known as tiles instead of the entire image as a whole. The adjacent tiles are then combined to remove any erroneous borders. This algorithm can be used by users to improve contrast in images. When only the brightness channel of an HSV image is altered, the effects are noticeably better than when all of the RGB image's channels are altered. Colored images can be processed using CLAHE, and the luminance channel is frequently employed for this (Ganesh & Ramesh, 2017).

#### 1.5.4.1 Parameters

Two most important parameters used while applying CLAHE are:

- **Clip limit**

  This variable controls the contrast limiting threshold. 40 is the default value.

- **Tile grid size**

  This determines how many tiles are in each row and each column. This defaults to being 88. It is utilized for applying CLAHE while the image is tiled.

### 1.5.5 Edge Detection

Edges are significant, regional changes in intensity in a digital image. It is a group of connected pixels that contains different regions. The three different edge types are horizontal, vertical, and diagonal. Edge detection is an algorithm for distinguishing areas of discontinuity in a picture. It is a frequently used technique in digital image processing, such as

- Pattern identification
- Morphology of images
- Extraction of features

Edge detection allows analysts to look at an image's properties for a significant change in the gray level. A new region in the image and the end of an existing one are indicated by this texture. While keeping the structural elements of an image, it reduces the amount of data in the image. Edge Detection Operators are of two types (Figure 1.6):

1. Gradient

First-order derivations are computed using gradient-based operators in digital images, such as the Sobel, Prewitt, and Robert operators.

2. Gaussians

An operator based on Gaussians that computes second-order derivations in digital images, such as the Canny edge detector and the Laplacian of Gaussian.

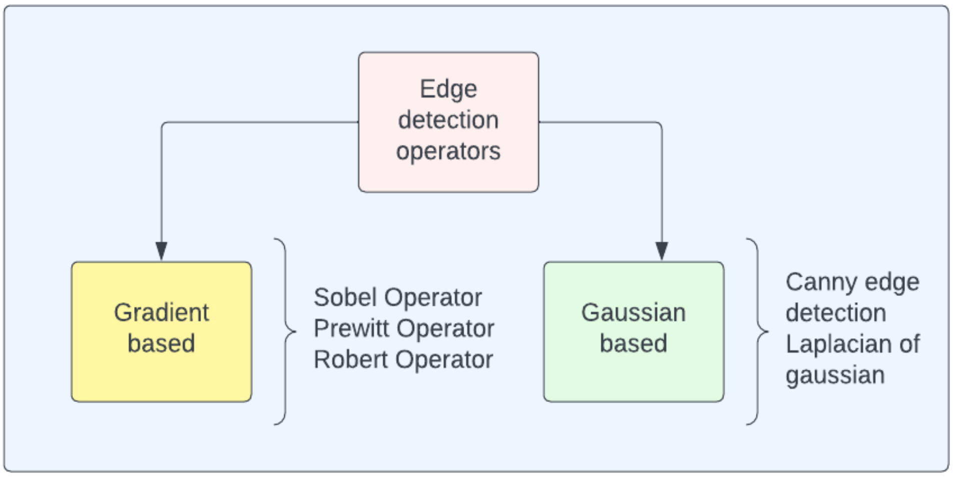


**Figure 1.6: Edge detection operator types.**

#### 1.5.5.1 Sobel Filter

The Sobel operator, also known as the Sobel-Feldman operator or Sobel filter, is a tool used in computer vision and image processing, particularly in edge detection techniques where it highlights edges in the resulting image. It bears the names of two colleagues from the Stanford Artificial Intelligence Laboratory, Irwin Sobel and Gary Feldman (SAIL). In a discussion at SAIL in 1968, Sobel and Feldman introduced the concept of a "Isotropic 3x3 Image Gradient Operator." The Sobel-Feldman operator uses a tiny, separable, integer-valued filter to convolve the image in both the horizontal and vertical axes. As a result, it is computationally reasonably cheap. However, it offers a somewhat rudimentary gradient approximation, especially for high-frequency fluctuations in the image.

##### 1.5.5.1.1 Advantages

The Sobel filter is simple and does really quick calculations. It is very proficient at looking for straight edges.

#### 1.5.5.1.2 Limitations

Sobel filters are quite sensitive to noise and give poor results when an image contains a lot of noise in it. If the images contain thick and uneven edges it does not detect them quite easily.

### 1.5.6 Stardist

From their website documentation: "Martin Weigert and Uwe Schmidt’s excellent deep learning based approach for 2D and 3D nucleus detection is called StarDist. It is available as a python library among other variations. A deep learning model that has been trained to recognise particular types of nuclei powers StarDist. To find nuclei in various types of photos, various models need to be trained. The general method for 2D images is shown in the following figure (Figure 1.7). Corresponding pairs of input (i.e raw) photos (A) and completely annotated label images make up the training date (i.e. every pixel is labeled with a unique object id or 0 for background). An overcomplete set of candidate polygons is generated for a given input image by training a model to densely predict the distances for a given input image by training a model to densely predict (D) the distance r (B) to the object border along a specific set of rays and object probabilities d (C). Through non-maximum suppression (NMS) of these candidates, the ultimate result is obtained (Weigert et al., 2020).

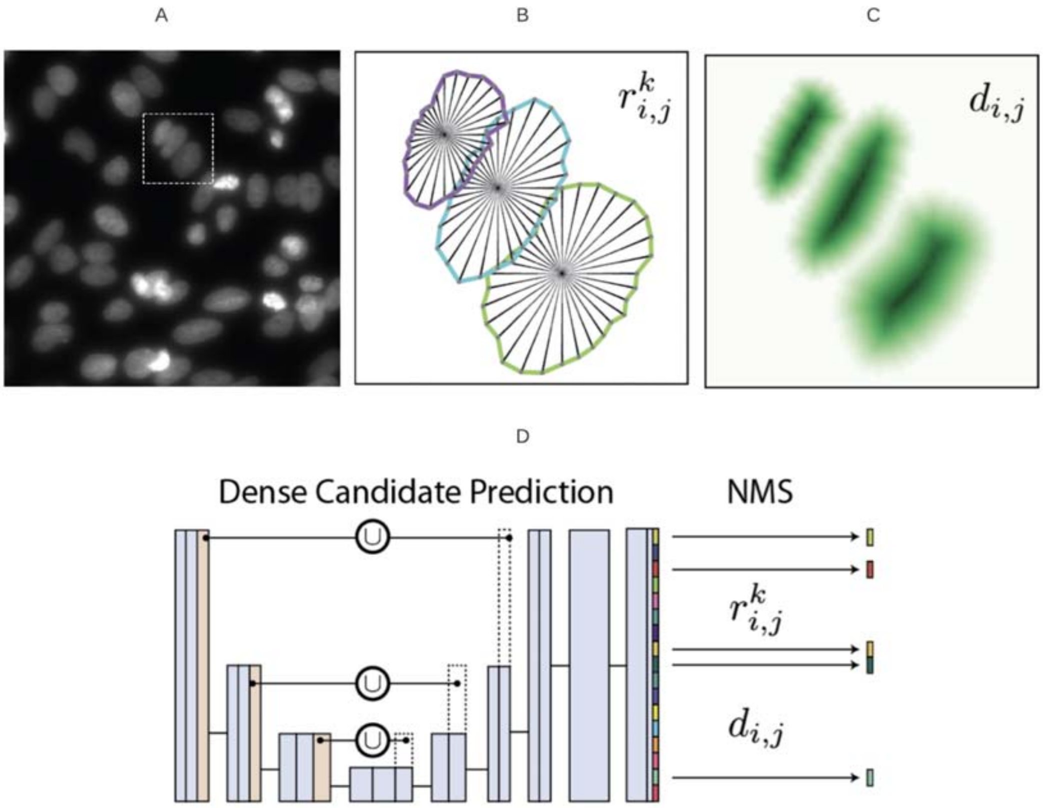


**Figure 1.7: Stardist edge detection process.**

## 1.6 Colon Cancer

The two leading causes of death in developed countries are colon and stomach cancers, with colon cancer coming in second for women and third for men in 2018 (Qaiser et al., 2019) (Iizuka et al., 2020). The architectural design of the gland's formation and its shape are used to diagnose colon ADC, the most popular forms (Iizuka et al., 2020). Segmentation can be used to localize information that is particular to the glandular regions, such as texture and cell arrangement in space. Many researchers used the MICCAI Gland Segmentation (GlaS) Challenge in 2015 by uploading photographs with annotations to show gland segmentation in their investigations (Kainz et al., 2017) (Iizuka et al., 2020).

### 1.6.1 Symptoms

Constant changes in bowel habits, such as gastroenteritis or constipation, or changes in stool consistency, as well as bleeding from the rectum or blood in the faeces, and ongoing abdominal pain, gas, or discomfort. Early on in the progression of the disease, many patients with colon cancer don't exhibit any symptoms. Depending on the size and location of the cancer, the signs and symptoms of a large intestine cancer can differ.

### 1.6.2 Causes

The majority of colon cancers, according to medical professionals, have unclear causes. Typically, Genetic changes in healthy colonic cells are what lead to the development of colon cancer (mutations). The DNA of a cell has a set of instructions that direct it in what to perform. For the body to continue to operate normally, healthy cells divide and expand in an orderly fashion. But, when a cell's DNA is harmed and it becomes cancerous, it continues to divide even when more cells are not required. A tumor is created as the cells assemble. The neighboring healthy tissue may eventually be destroyed when the cancer cells spread and eat through the area. Furthermore, malignant cells may move to various body regions and establish themselves there (metastasis).

### 1.6.3 Screening colon cancer

Doctors urge those with an average risk of colon cancer to consider getting checked around the age of 45. Yet individuals who are more at risk, such those with a family history of colon cancer, need to start planning their screenings earlier.

## 1.7 Feature extraction

A step in the process that aims to eliminate extraneous or duplicate information is feature extraction. An initial set of raw data is separated and summarized during this procedure, and data items that are unimportant to the task at hand are removed. Consequently it will be simpler once processing is complete. The presence of a huge number of variables in large datasets is their most crucial feature. Processing these variables takes a lot of computing power. In order to efficiently reduce the amount of data, feature extraction helps to extract the best features from those large datasets by choosing and combining variables into features. This subset of features is simple to handle while still accurately and uniquely describing the actual dataset.

## 1.8 K-NN

The K-Nearest Neighbors algorithm (K-NN) in statistics was created by Evelyn Fix and Joseph Hodges in 1951 and later improved by Thomas Cover. It is a non-parametric supervised learning technique. Regression and classification are two uses for it. The input in both situations consists of a data set's k closest training samples. Whether K-NN is applied for regression or classification determines the results.

### 1.8.1 K-NN classification

A class membership comes from the K-NN classification process. The class to which an item is assigned depends on the majority vote of its neighbors, who then assign it based on the preferences of the object's k closest neighbors (k is a positive integer, typically small). If $k = 1$, the object is simply assigned to the class of its one closest neighbor.

### 1.8.2 K-NN regression

The object's property value is the result of K-NN regression. The k nearest neighbors average is represented by this number. If $k = 1$, only the value of the output's one nearest neighbor is assigned. With K-NN, the function is just locally approximated and all computation is delayed until after the function has been assessed. While this method uses distance to classify objects, normalizing the training data can greatly improve accuracy if the features represent a variety of physical units or have extreme size variations. Giving neighbors' contributions weights, with the closest neighbors contributing more to the average than the distant neighbors, is a useful technique for categorization and regression. The neighbors are chosen from a set of objects for which the class or object property value is known when using K-NN classification or regression. This can be regarded as the algorithm's training set even though there is no requirement for a specific training step. The K-NN methodology has the ability to be sensitive to the data local structure.

### 1.8.2 Algorithm

During the training phase of the method, just the extracted features and corresponding labels of the training set are saved. During the classification phase, a query point is given the label that appears the most often among the k training samples that are the closest to the unprocessed vector (a test point), where k is a user-defined constant. A common distance metric for continuous variables is euclidean distance. By applying specific techniques, such as Neighbour Component Analysis, to learn the distance metric, it is typically capable of significantly enhancing the accuracy of K-NN.

### 1.8.3 Parameter selection

The optimum value of k relies on the information currently available; usually, greater values of k reduce the precision of classification distinctions while minimizing the effect of noise on classification. There are numerous computational techniques that can be used to select an appropriate k.

### 1.8.4 Validation of results

It is frequently done using a confusion matrix, also known as a "matching matrix," to check the precision of K-NN classification.

### 1.8.5 Advantages of K-NN

i) K-NN is quite slow in learning the features. It requires no training during the training time. It only tends to make use of the training dataset that is kept and learns from it while delivering real-time predictions. K-NN is hence quicker.

ii) Because the K-NN method does not require training before predicting things, additional data can be added without impacting the reliability of the system.

iii) Applying K-NN is relatively easier than other machine learning algorithms. The parameter K and the distance computing function, such as the Manhattan or Euclidean distance, are the sole two parameters required for K-NN implementation.

### 1.8.6 Disadvantages

i) Because it is extraordinarily expensive to figure out the distance between each point and each change in the number in a large dataset, K-NN does not perform well with massive datasets.

ii) Because calculating distances in each dimension becomes increasingly difficult as the number of dimensions rises, the K-NN algorithm performs badly with huge dimensional data.

iii) K-NN is sensitive to dataset noise so manually imputing missing numbers and eliminating outliers is required.

## 1.9 Nature of problem

The amount of medical imaging data is growing, as is the need for quick computers to analyze such complicated images. Images of colon cancer from histology are used in this study's data. Finding patterns in histopathology slide images that may be used to categorize them and determine whether they are benign or malignant.

## 1.10 Proposed solution

The aim of this study is to build a software that could help in identifying cancer in images (histopathology slide). This is done by first collecting data of histopathology images and then finding parameters that can help in categorizing it. Once these parameters are extracted they are fed to a machine learning model to train it. Subsequently new histopathology images may be provided to the software that allows it to categorize the image as having benign or malignant tumor based on the machine learning model. Chapter 2 (Materials and Methods) gives an overall methodology of the thesis and the tools used to achieve the objective. In Chapter 3, results are shown after applying strategies of the proposed model and performance of the proposed model through graphs and finally the conclusion stating the overall importance of this thesis.

The ultimate solution is to first use fast computers that will help in analyzing the whole slide histopathology images of colon cancer dataset. After gathering image data the next step is to adjust contrast and brightness of images using the normalization function of OpenCV. Once image detail is increased, the next phase is to extract only the nuclear regions in the image as it contains meaningful information for analysis purposes. This is done by selective color segmentation. Edges of the nucleus are extracted and then bounding boxes are created around them for calculation of feature vectors. Once feature vectors are extracted then an algorithm for example K-Nearest Neighbor (K-NN) is applied to classify data as benign or malignant and also calculate overall accuracy of the model.

# MATERIALS AND METHODS

# 2. MATERIALS AND METHODS

This study is based on finding the visual patterns present in images from the colon dataset. Once the patterns have been recognised the next task is to quantise them statistically. When these patterns have been converted to numerical values the range of these values and their combinations are further classified as belonging to a benign or malignant class. A number of steps (Figure 2.1) were performed to find better results and to differentiate between the benign and malignant histopathology images.

Hematoxylin and eosin stained images are used to find the patterns present in these histopathology images. Images of a standard size were used. Brightness and contrast normalization was performed to improve the quality of the images. After image normalization, selective color extraction of ROI (region of interest) is applied. Nuclear regions with bright blue color were extracted by applying color segmentation technique which leaves behind only the ROIs in the image and removes any other visual information. Then contours (shapes) present in the remaining image were extracted. The contours are then converted to their bounding boxes and lastly these bounding boxes are rotated upright for appropriate calculations of the nuclear regions. Then a feature vector is created containing the nucleus area mean, nucleus area standard deviation, nucleus height mean, nucleus width mean and aspect ratio for each whole slide image based on these bounding boxes. Feature vectors for all images selected from the dataset are similarly extracted and then written to a CSV (comma-separated values) file to be used later. This file is used as the source of data for the K-NN model. The K-NN, does not require parameters and is a supervised machine learning categorization algorithm which depends on closeness to classify or predict how a specific data item would be categorized. A classifier is chosen for categorization issues by a simple majority vote, indicating that the term that is most commonly expressed around an unique data point is chosen. The model can then be provided with a new data item and queried if it is likely to fall into one of the specified classes or labels.

## 2.1 Workflow

This study used histopathology images of colon cancer dataset.

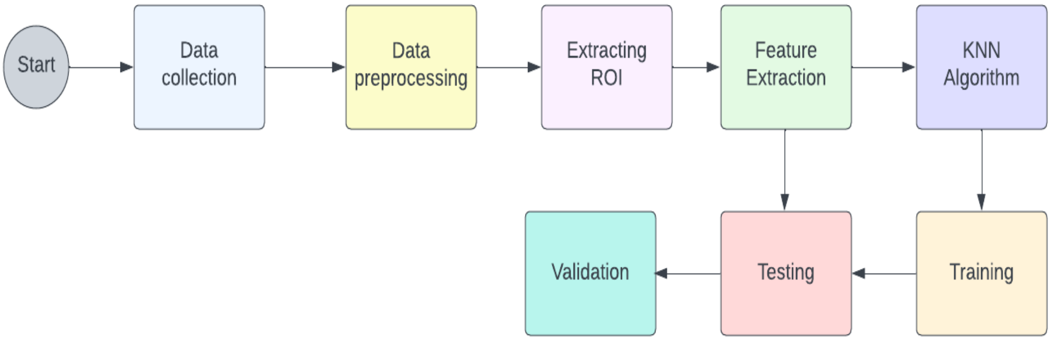


**Figure 2.1: Software workflow.**

## 2.2 Data collection

In this work, the K-NN algorithm is applied to the colon cancer dataset to classify tissue images as benign or malignant based on their nucleus properties present in histopathology images. This dataset was sourced from the Kaggle database. The dataset contains over 25,000 images of which 1090 colon tissue images were used in JPEG format including benign and malignant slides. This study used images containing the most observable differences between the two. K-NN model was applied to find the accuracy of the developed pipeline.

## 2.3 Data preprocessing

Data preparation is a crucial stage in both image analysis and machine learning. By removing undesirable distortions or increasing particular features that are essential for further processing, preprocessing seeks to enhance the image data. By enhancing the contrast and brightness of an image, these processes not only make it easier to see the image clearly but also make it easier to extract features from the image.

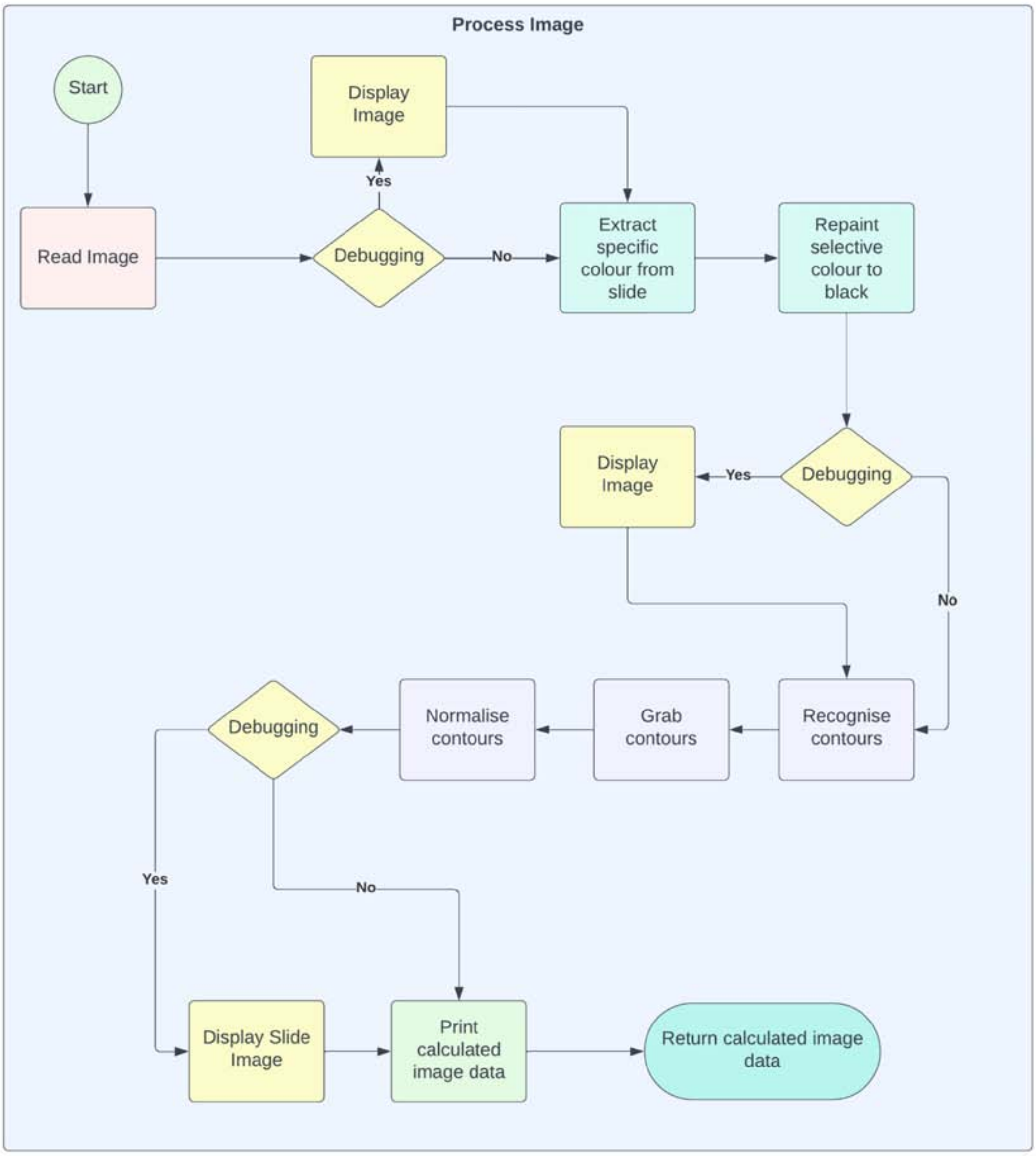


**Figure 2.2: Data preprocessing workflow.**

### 2.3.1 Image normalization

Image normalization is used to increase the brightness or contrast in an image. Brightness is a term used to describe the overall lightness or darkness of an image. After applying the normalization function of python the nuclear regions and the background has become quite clear. This step helped to properly visualize the overall image and also to further process the image.

## 2.4 Extracting ROI

The region of interest (ROI) is the nuclear region indicating the bright blue color in a histopathology image. So the color segmentation extracts regions out of an image that belong to a certain color range. This method separates an image into coloured parts by correlating each pixel's color feature to the color features of adjacent pixels or even a learnt color classifier. A specific range of color is selected to specify and extract the ROI (nuclei regions).

### 2.4.1 Contour detection

The contour function is used to identify the nucleus boundaries. Simply a contour is a curve that connects a collection of dots that enclose a region with the same color or intensity. The curve encircling this region is the contour denoting the shape of the object, and the area is of uniform color or intensity making an object that is necessary to detect. Consequently, contour detection operates in a manner similar to edge detection, with the exception that the edges must be in the shape of a closed path.

### 2.4.2 Stardist

Stardist function is used to distinguish among different nuclei instances. This helps to extract the feature vector values and to create a library of each image value in a .csv file. The nucleus bounding boxes are created after applying Stardist function. These bounding boxes are then further rotated upright for better calculation of results.

## 2.5 Feature extraction

One of the key steps is feature extraction. A list of characteristics or image features that will most successfully or meaningfully represent the information needed for evaluation and categorization are defined in this approach. An important field of research involves identifying the kind of cancer from the patterns in histopathological images. The numerous characteristics found in this research's data include;

### 2.5.1 Nucleus mean area

Nucleus mean area is calculated for each of the histopathological images of the colon cancer dataset. This feature is used to calculate the overall mean area in each image. This feature helps to find the overall average area occupied by the nucleus in a histopathology image.

### 2.5.2 Nucleus area standard deviation

Nucleus area standard deviation is used to calculate the average amount of variability in the nucleus area. This nucleus area standard deviation helps to find the smallest and the largest nucleus present in an image.

### 2.5.3 Nucleus mean height

The average height of the nuclei visible in each image of the dataset is determined by calculating the nucleus mean height of every nucleus in an image. It is useful to recognise the general pattern in the nuclei's height.

### 2.5.4 Nucleus mean width

The average width of the nuclei visible in each image of the dataset are determined by finding the nucleus' mean width. It is useful to recognise the general pattern in the nuclei's width to find the general trend of width in the benign and malignant images.

### 2.5.5 Aspect ratio

Aspect ratio is used to calculate a proportional relationship between an image's nucleus widths and heights mean. If the value is 1 it means the nucleus present in the image are circular in shape as widths and heights are the same but if the value is greater than 1 it means the nucleus are not circular and are deformed.

### 2.5.6 Malignant

This feature is used to show the type of image as benign or malignant. 0 is for the benign and 1 is for the malignant. This is only available in the training and test dataset.

### 2.5.7 Correlation

A modified probability is used by the correlation-based K-Nearest-Neighbor algorithm to speed up calculation and improve prediction accuracy and classifies data based on the correlation calculation (Li & Xiang, 2012). The correlation was performed between the feature vectors and malignancy. This correlation is used to find the effect of feature vectors on malignancy. This correlation technique is also used to find the feature which has the most effect on the malignancy.

## 2.6 K-NN

A supervised learning method called K-Nearest Neighbour (Eyheramendy et al., 2003) classifies the results of new instance queries using the K-Nearest Neighbour categories scale. The key to the K-NN algorithm is to find the closest neighbors. Vectors with class labels in a multidimensional feature space make up the training samples. Only the feature vectors and class labels of the training samples are stored during the algorithm's training phase. K is a user-defined constant used in the classification phase, and an unlabeled vector (a query or test point) is classified by assigning the label that appears most frequently in the K training samples that are closest to the query point (Muzaffar Khan et al., 2011). Different values of K are used to find an instance where the accuracy of the model is quite higher than others (Muzaffar Khan et al., 2011).

### 2.6.1 Training and Test dataset

In a typical machine learning model, the original dataset is divided into two parts. The first set is known as training data - which is a fraction of the original data that is then analyzed into our model to learn the patterns, which in turn trains our model. In our case our training set contained 870 images, which was 80% of the total dataset. Usually the data in the training phase is relatively more than the training phase. The machine learning model requires as much data as possible so that the model may recognize the underlying pattern. Training set varies greatly according to the type of machine learning we use: supervised or unsupervised, while the other subset is known as the test data. Test data is unseen data to test our machine learning model. This study used 220 images as 20% of the total dataset as a test data.

### 2.6.2 Validation

The model validation is used to find out how accurate a model is performing classification. Accuracy is the measure of correct predictions generated for the test dataset. This is done by feeding the model with a test dataset. It is calculated once the correct predictions are calculated. After finding the correct predictions its percentage is calculated from the total predictions finding the true positives and the true negatives in the test data.

## 2.7 Software flow diagram

### 2.7.1 Main software loop

#### 2.7.1.1 Feature vector extraction

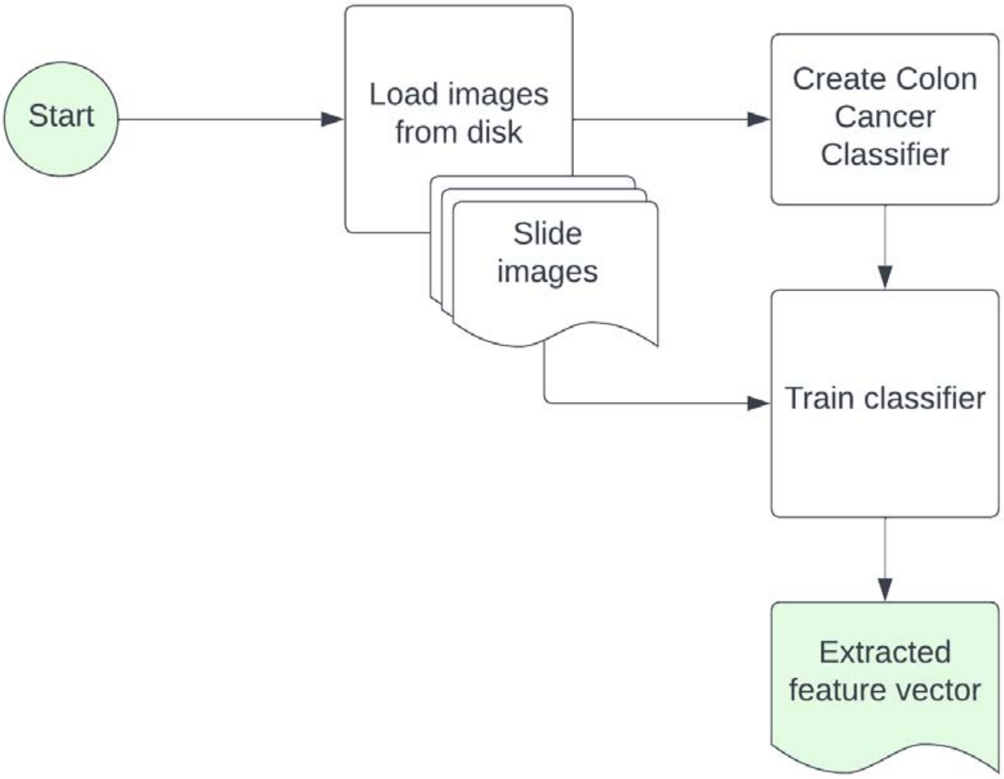


**Figure 2.3: Feature vector extraction.** Feature vector is extracted by loading the dataset images first and then creating a colon cancer classifier object that implements extraction logic. These images are passed to the classifier and it processes each image in turn, returning the final extracted feature vector.

### 2.7.1.2 K-NN model training and validation

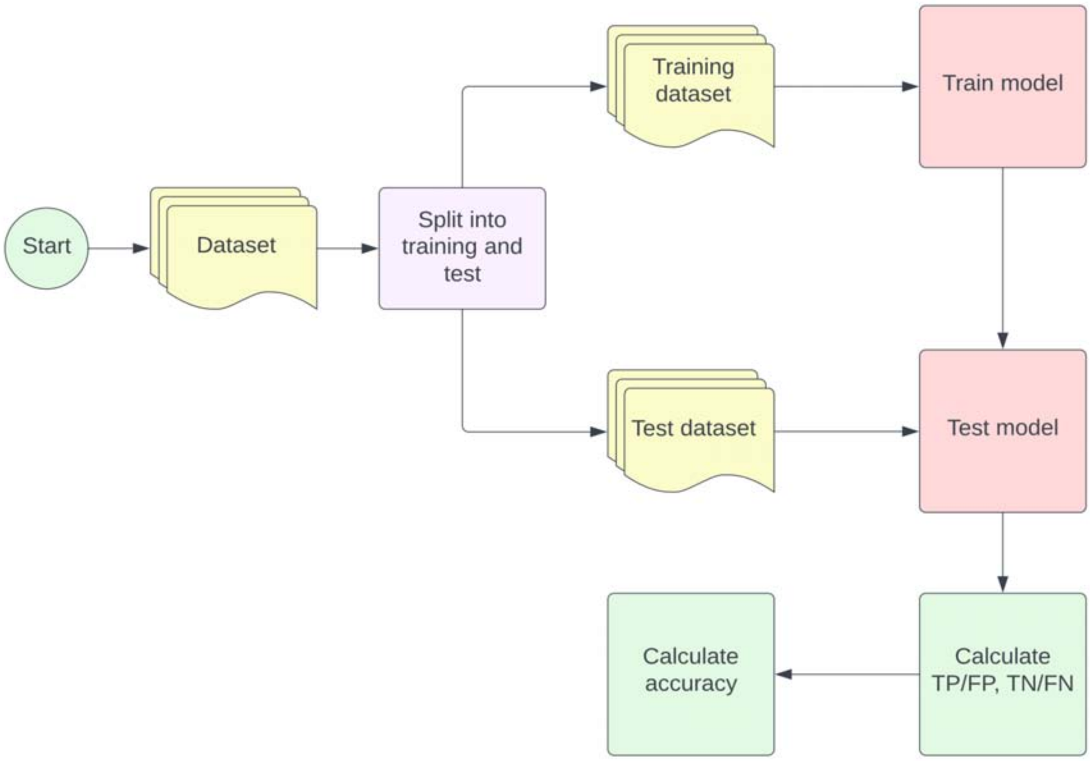


**Figure 2.4: K-NN model training and validation.** The K-NN model is trained by loading the feature vector extracted in an earlier part of the software, splitting it into the training and test dataset. The training dataset is then fed to the blank model to allow it to identify the vector. Then the model is given the test dataset and it makes predictions against that test dataset. Afterwards the known malignancy values of the test dataset are compared against the predictions generated by the model to calculate the true positive, false positive, true negative and false negative scores. From these data points the accuracy of the model can be further calculated.

### 2.7.2 Libraries used

The libraries used in this software are:

- Matplotlib (3.5.0)
- Numpy (1.23.3)

- OpenCV (4.6.0.66)
- Scikit (1.2.1)
- Stardist (0.8.3)
- Pandas (1.5.1)

# RESULTS

# 3. RESULTS

This section provides the results obtained by implementing the already described methodology.

## 3.1 Image normalization

Image normalization is used to increase the brightness or contrast in an image. Brightness is a term used to describe the overall lightness or darkness of an image (Figure 3.1). After applying the normalization function of OpenCV to the original slide image (Figure 3.1) the nuclear regions and the background have become quite clear (Figure 3.2) in the normalized slide image. This step helped to properly visualize the overall image and also to further process the image.

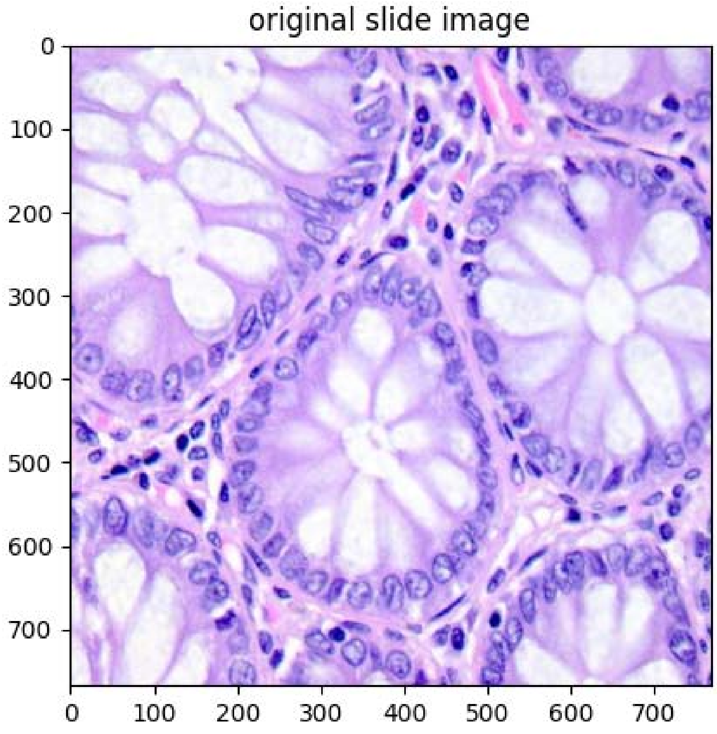


**Figure 3.1: Original histopathology slide image.** The white part in the center of the flower distribution is the lumen. The other light purple and white parts surrounding the lumen are the epithelial cells. The epithelial cell distribution is surrounded by the dark blue, purple nuclei. Outside the nuclei ring are the dark blue stroma.

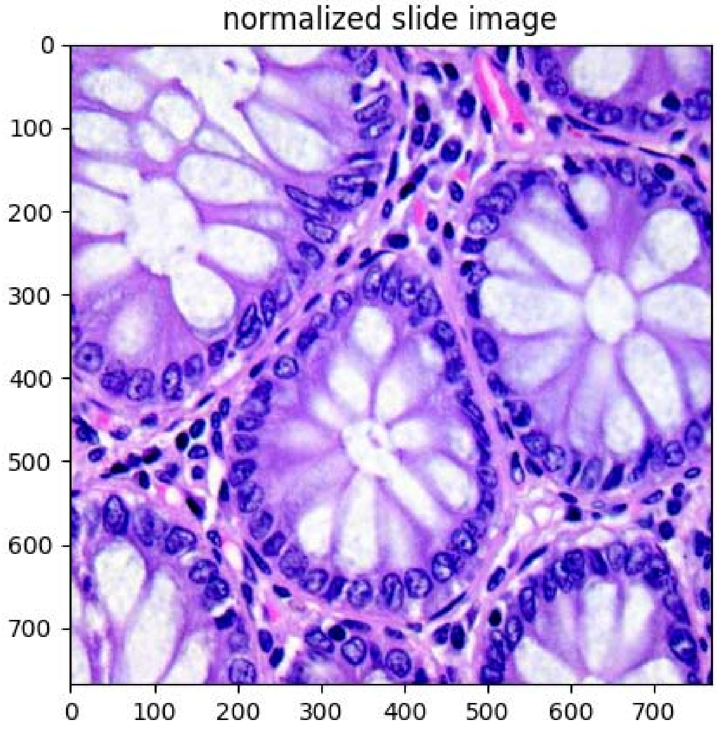


**Figure 3.2: Normalized slide image.** The nuclei ring and stroma are darkened and more prominent in the normalized slide image.

## 3.2 Extracting ROI

The region of interest is the nuclear region indicating the bright blue color in a histopathology image. So the color segmentation extracts regions out of an image that belong to a certain color range. This method divides an image into coloured regions by comparing each pixel's color feature to the color values of adjacent pixels or a trained color classifier. A specific range of color is selected to specify and extract the ROI (nuclei regions) (Figure 3.3).

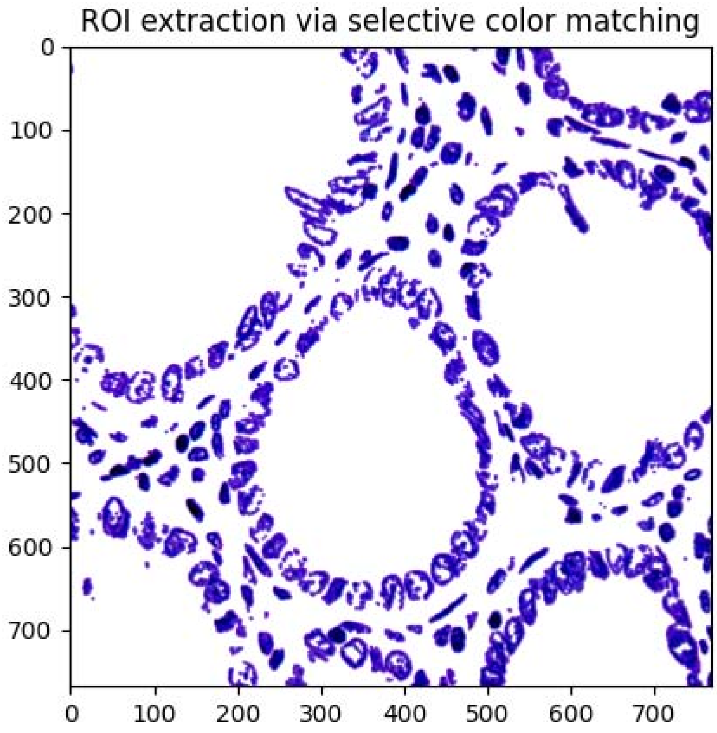


**Figure 3.3: Selective color segmentation.** Color range extraction leaves only the regions of interest in the image while removing all other parts.

### 3.2.1 Contour detection

The contour function is used to identify the nucleus boundaries. Simply a contour is a curve that connects a collection of dots that enclose a region with the same color or intensity. The curve encircling this region is the contour denoting the shape of the object, and the area is of uniform color or intensity making an object that is necessary to detect. Consequently, contour detection operates in a manner similar to edge detection, with the exception that the edges must be in the shape of a closed path (Figure 3.4).

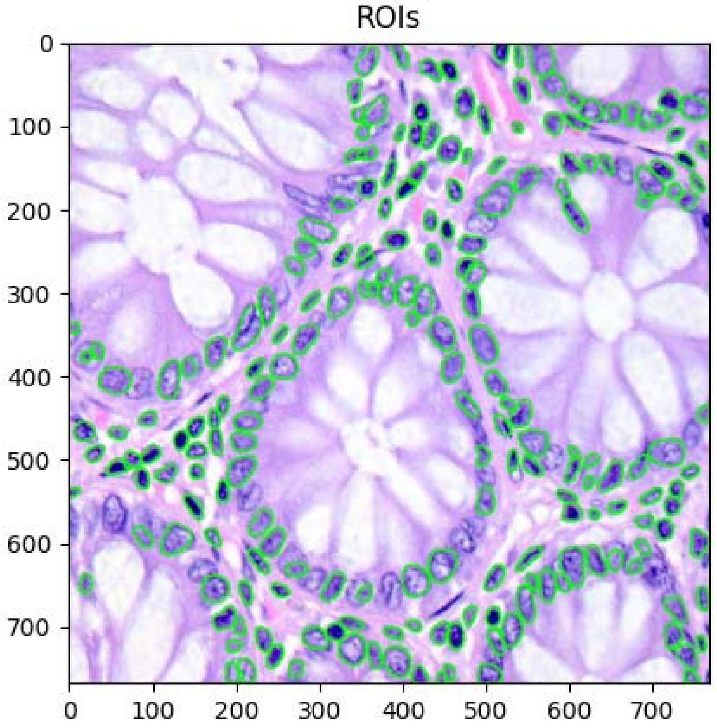


**Figure 3.4: ROI contour detection.** The green outline denotes the contour or shapes of ROI that are detected by the algorithm.

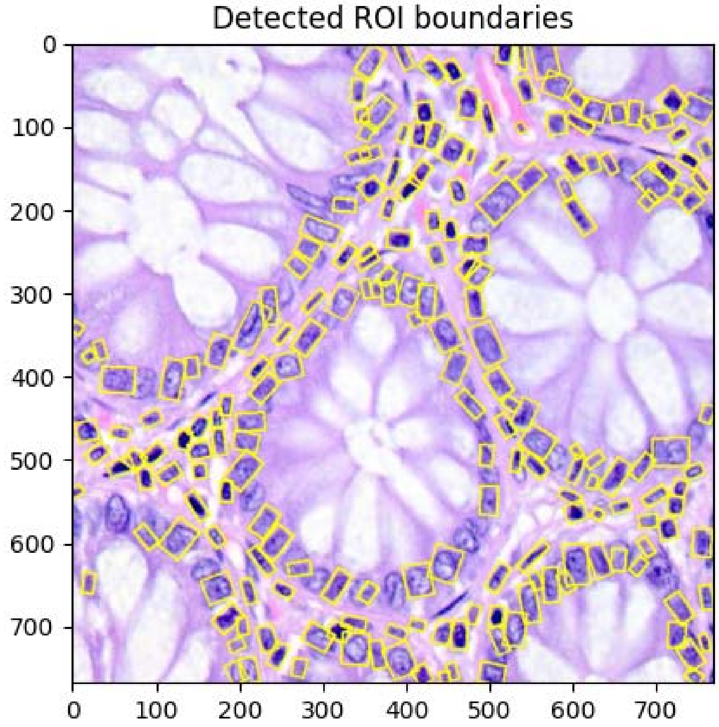


**Figure 3.5: ROI contour to bounding box conversion.** Yellow boxes show rectangular boundaries of the ROI that were detected.

### 3.2.2 Stardist

Stardist function is used to distinguish among different nuclei instances. This helps to extract the feature vector values and to create a library of each image value in a .csv file. The nucleus bounding boxes are created (Figure 3.5) after applying the Stardist function. These bounding boxes are then further rotated upright for better calculation of results (Figure 3.6).

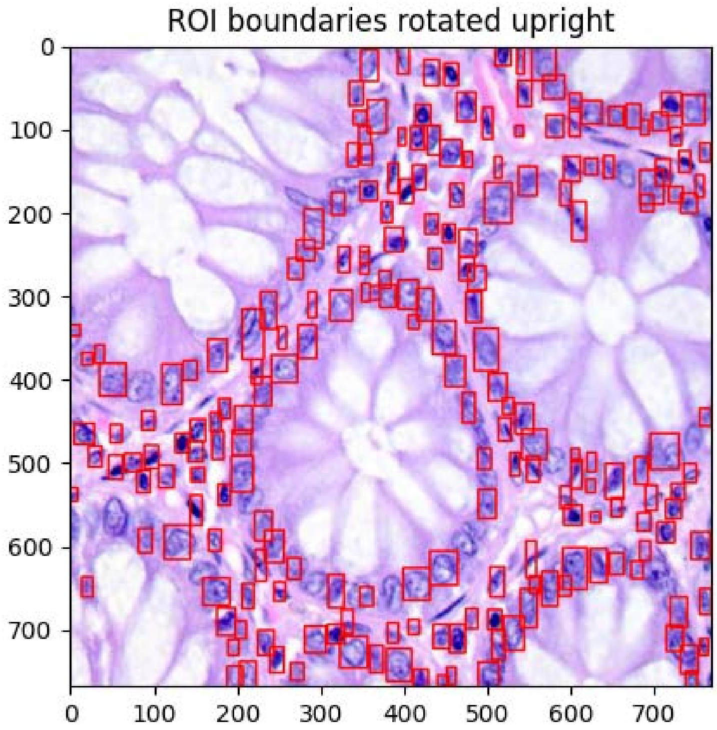


**Figure 3.6: ROI bounding box rotation.** Red boxes show the boundary rectangles around the detected ROI that have been rotated upright for further easier calculation.

## 3.3 Feature extraction

One of the necessary phases is feature extraction. A set of features or image characteristics that will most effectively or functionally reflect the data required for analysis and classification are defined in this approach. An important field of research means determining the kind of cancer from the patterns in histopathological images. The various features (Figure 3.7) present in this research data are;

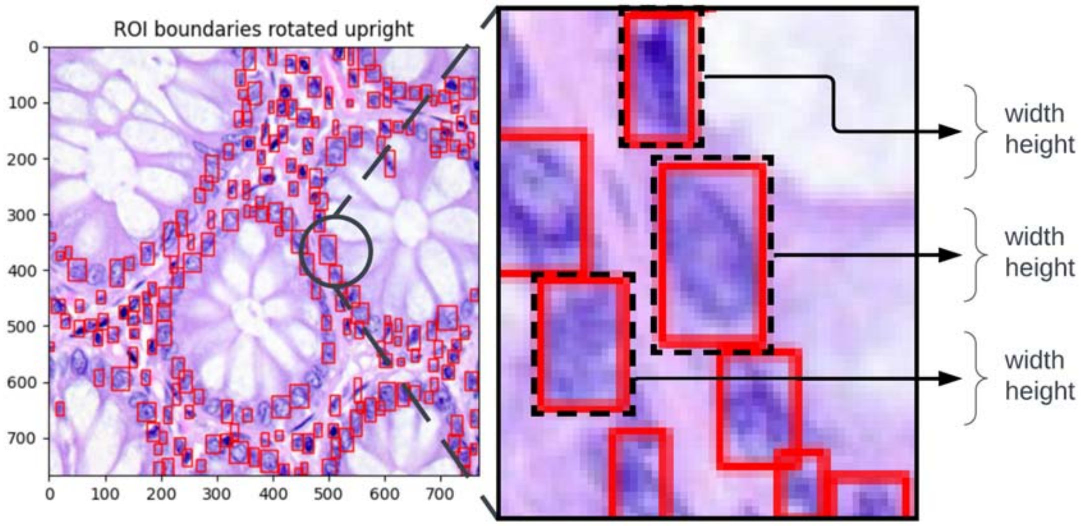


**Figure 3.7: Value extraction from ROI bounding box.** Red boxes show each detected ROI and the width and height values that have been calculated from these regions.

### 3.3.1 Nucleus mean area

Nucleus mean area is calculated for each of the images to calculate the overall mean area in each image of the dataset to find the overall average area occupied by the nucleus in a histopathology image. From the graph of the nucleus mean area it can be seen that the malignant mean area is quite high from around 300 to 900 whereas the benign nucleus mean area is from 230 to 280. Only a few nucleus mean areas are in the benign region but the majority mean area is quite higher.

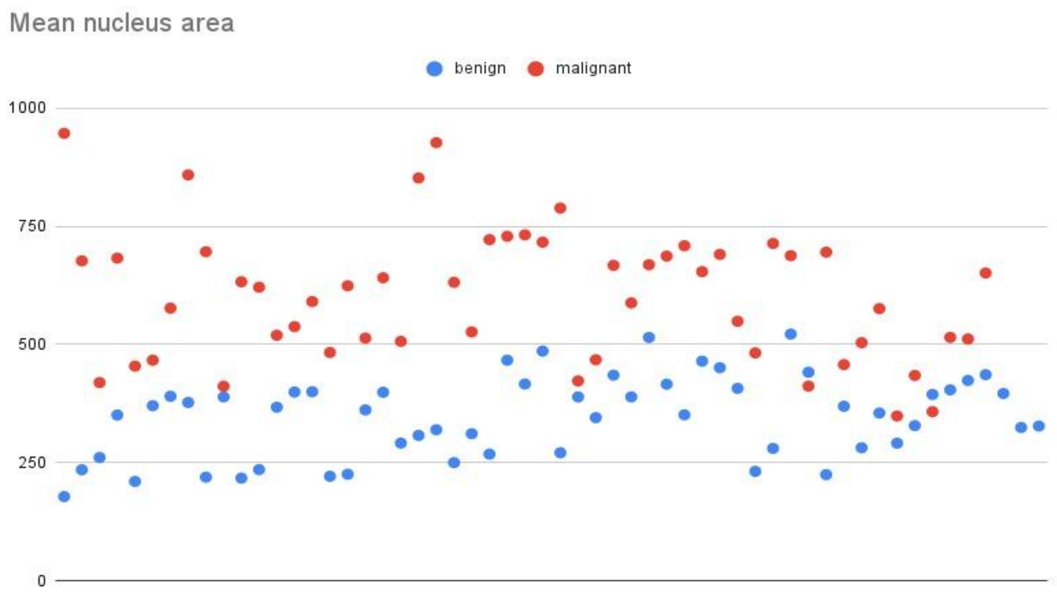


**Figure 3.8: Mean nucleus area distribution in training dataset.** This graph shows the distribution of mean nucleus area for each histopathology image. The blue points are for benign slide images while the red points show mean nucleus area for malignant slide images.

### 3.3.2 Nucleus area standard deviation

Nucleus area standard deviation is used to calculate the average amount of variability in the nucleus area.

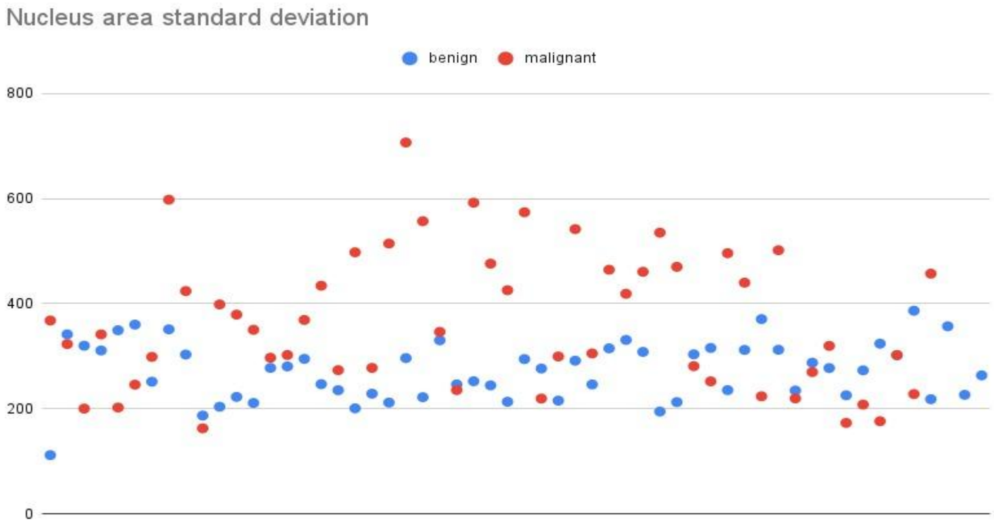


**Figure 3.9: Nucleus area standard deviation distribution in training dataset.** This graph shows the distribution of nucleus area standard deviation for each slide image, benign and malignant slide images shown in blue and red, respectively.

### 3.3.3 Nucleus mean height

The average height of the nuclei visible in each image of the dataset are determined using the nucleus mean height. It is useful to recognise the general pattern in the nuclei's height. The nucleus mean height is quite higher in the malignant nucleus whereas the mean nucleus height in the benign is relatively lower which shows that the benign nucleus height is less than the malignant nucleus whereas the malignant nucleus are quite elongated and deformed in shape.

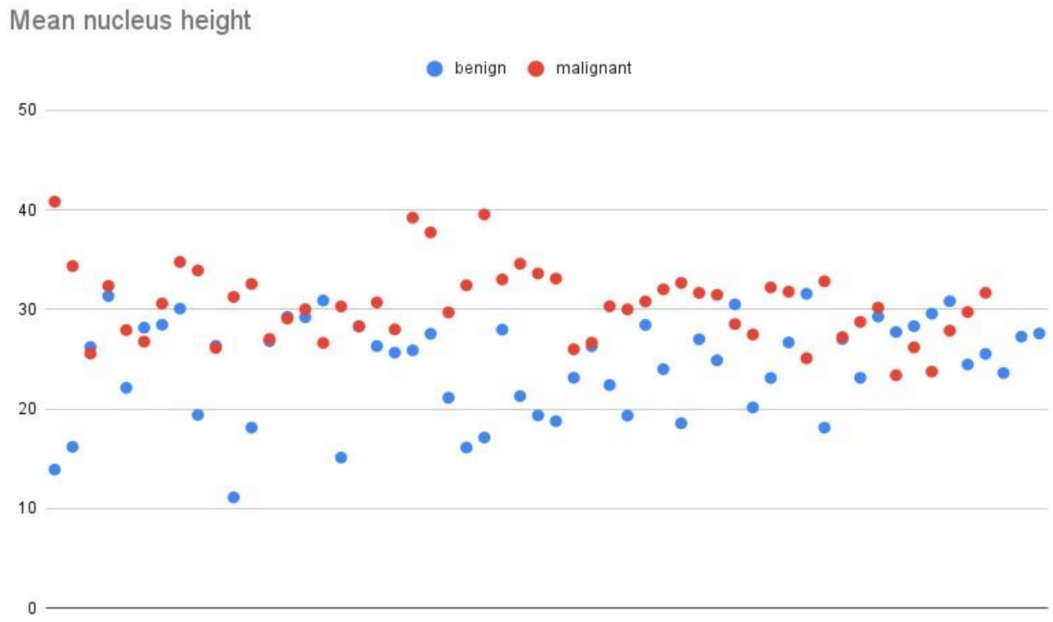


**Figure 3.10: Mean nucleus height distribution in training dataset.** This graph shows the mean nucleus height for each slide image where the blue dots show values for the benign images while the red dots depict malignant slide images.

### 3.3.4 Nucleus mean width

The average width of the nuclei visible in each image of the dataset are determined using the nucleus mean width. It is useful to recognise the general pattern in the nuclei's width. The mean nucleus width values for both benign and malignant nucleus are almost same which means that the malignant nucleus don't grow horizontally but increase vertically and are quite stretched.

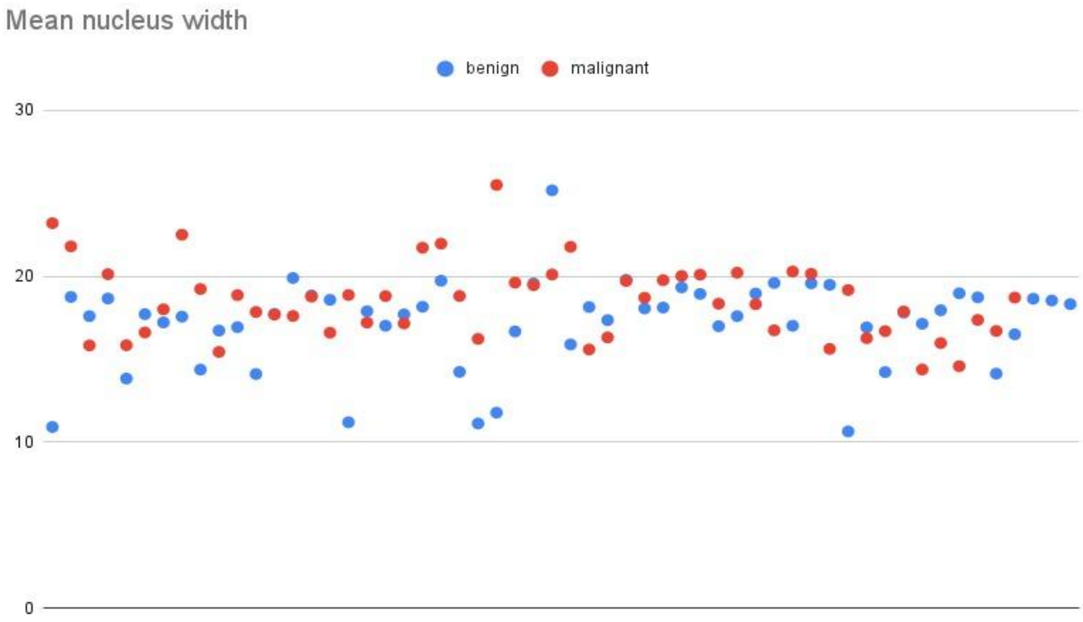


**Figure 3.11: Mean nucleus width distribution in training dataset.** This graph shows the mean nucleus width for each slide image. The blue series is benign slide images while the red series is malignant slide images.

### 3.3.5 Aspect ratio

Aspect ratio is used to calculate a proportional relationship between an image's nucleus widths and heights mean. If the value is 1 it means the nucleus present in the image are circular in shape as widths and heights are the same but if the value is greater than 1 it means the nucleus are not circular and are deformed.

### 3.3.6 Malignant

This feature is used to show the type of image as benign or malignant. 0 is for the benign and 1 is for the malignant. This is only available in the training and test dataset.

### 3.3.7 Correlation

The correlation was performed between the feature vectors and malignancy. ROI area mean and ROI aspect ratio gives the highest correlation of 0.748694 and 0.620070 respectively whereas ROI width mean and ROI area std dev gives the lowest correlation with malignancy. As it was observed through their results that theROI width mean and ROI area std dev do not provide a good correlation to malignancy so they were dropped before the training phase.

**Table 3.1: Feature correlation matrix.**

| | |
|---|---|
| **ROI area mean** | 0.748694 |
| **ROI area std dev** | 0.433807 |
| **ROI width mean** | 0.256647 |
| **ROI height mean** | 0.571090 |
| **ROI aspect ratio mean** | 0.620070 |

## 3.4 K-NN

The K-NN is used to categorize the dataset as benign and malignant. The value of k is set to 8 as accuracy of the model is very high at this point as can be seen from the graph given below. The peak is relatively higher than other values at k=8 which clearly shows that this point is quite suitable for the current model. Specifically accuracy at k=8 is 90.91%.

## 3.5 Data distribution

In the splitting step, the data is divided into two or more subsets. Normally, in a two-part split, one part of the dataset is used to train the model and another part is used to test or evaluate the data. It is a common practice to divide the data into 80% training and 20% as a testing set. A similar technique was followed and the dataset was split into 870 images as the training set and 220 images as the test set. Training set is used to estimate different parameters and the testing set is used after the training is done to find the accuracy of the model and how accurately the model is predicting.

A few feature vectors from the resulting dataset are recorded below for example.

**Table 3.2: Samples from training dataset feature vector.**

| ID | area_mean | area_std | width_mean | height_mean | aspect_ratio_mean | malignant |
|---|---|---|---|---|---|---|
| 1 | 178.120938 | 111.72835 | 10.912389 | 13.912873 | 1.252357 | 0 |
| 2 | 234.876853 | 341.63494 | 18.748503 | 16.192387 | 1.529571 | 0 |
| 3 | 260.786868 | 319.96290 | 17.590643 | 26.216374 | 1.527601 | 0 |
| 4 | 350.643535 | 310.9163 | 18.656051 | 31.341263 | 1.702084 | 0 |
| 5 | 210.123579 | 349.37893 | 13.830129 | 22.123089 | 1.602733 | 0 |
| 6 | 370.431231 | 360.12847 | 17.710526 | 28.164474 | 1.553522 | 0 |
| 7 | 390.343243 | 251.53446 | 17.215859 | 28.458155 | 1.602116 | 0 |
| 8 | 377.232133 | 351.12544 | 17.562092 | 30.078431 | 1.663183 | 0 |
| 9 | 219.242142 | 303.16197 | 14.372103 | 19.398532 | 1.627223 | 0 |
| 10 | 389.132132 | 187.07394 | 16.722408 | 26.337793 | 1.601838 | 0 |

### 3.5.1 Training and test dataset

In a typical machine learning model, the original dataset is divided into two subsets. The first set is known as training data - which is a fraction of the original data that is then analyzed into our model to learn the patterns, which in turn trains our model. In our case our training set contained 870 images, which was 80% of the total dataset. Training set is usually larger than the test set. The machine learning model requires as much data as possible so that the model may recognize the underlying pattern. Training set varies greatly according to the type of machine learning we use: supervised or unsupervised, while the other subset is known as the test data. Test data is unseen data to test our machine learning model. This study used 220 images as 20% of the total dataset as a test data.

## 3.6 Accuracy

Accuracy is the measure of correct predictions generated for the test dataset. The total accuracy for the current model is 90.91%. This is because out of 220 images of the test data only 20 images were inaccurately predicted.

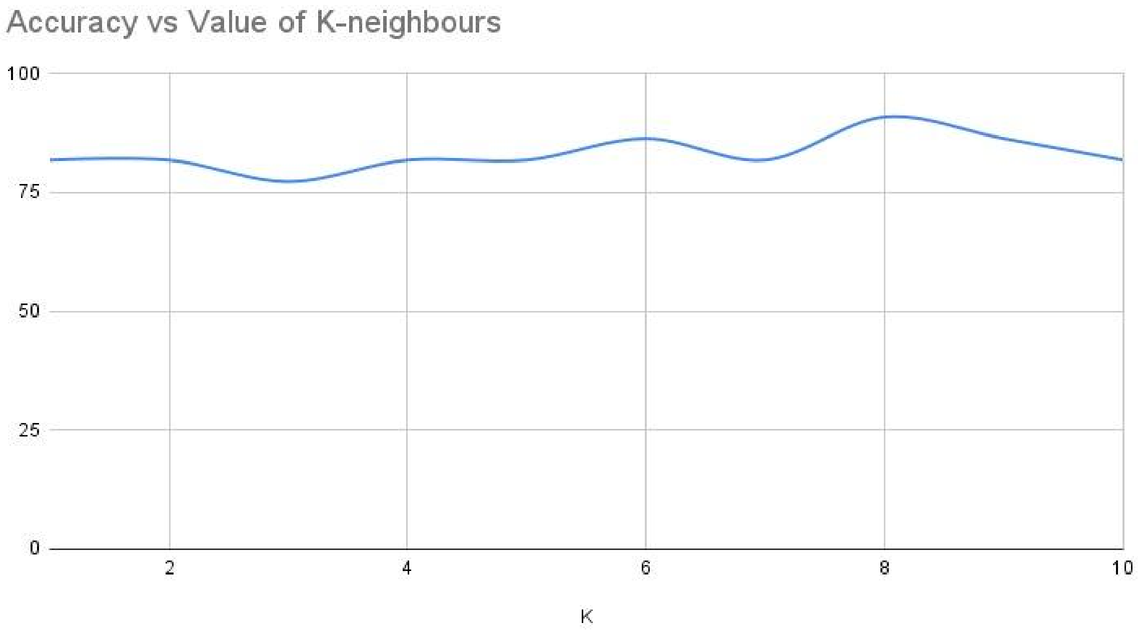


**Figure 3.12: K-NN model accuracy vs value of K-neighbors.** The graph shows that as we change the value of K-neighbors the accuracy of the model changes. For most values of k it hovers around 80% whereas the peak is achieved at k=8 at 90.91%.

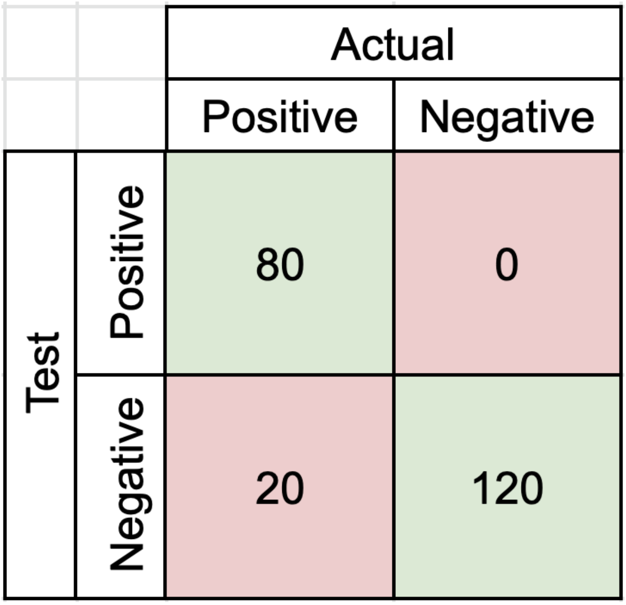
**Table 3.3: K-NN model accuracy as value of k changes.**

| K | Num validations | Errors | Error rate Percentage | Accuracy Percentage |
|---|---|---|---|---|
| 1 | 220 | 40 | 18.18 | 81.81 |
| 2 | 220 | 40 | 18.18 | 81.81 |
| 3 | 220 | 50 | 22.72 | 77.27 |
| 4 | 220 | 40 | 18.18 | 81.81 |
| 5 | 220 | 40 | 18.18 | 81.81 |
| 6 | 220 | 30 | 13.63 | 86.36 |

| | | | | |
|---|---|---|---|---|
| 7 | 220 | 40 | 18.18 | 81.81 |
| 8 | 220 | 20 | 9.09 | 90.91 |
| 9 | 220 | 30 | 13.63 | 86.36 |
| 10 | 220 | 40 | 18.18 | 81.81 |

### 3.6.1 Validation

The accuracy of 90.91% is obtained. Accuracy is used to find the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.



| | | Actual | |
|---|---|---|---|
| | | Positive | Negative |
| Test | Positive | 80 | 0 |
| | Negative | 20 | 120 |

**Figure 3.13: Confusion matrix for k=8.**

(Figure 3.13) shows these values for k=8. Out of a total 220, true positives are 80, true negatives are 120. False positives are 0 and false negatives are 20. This results in a sensitivity of 0.80 and a specificity of 1.0. False negative rate is 0.20 whereas false positive rate is 0.0.

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots true positive and false positive rates. The true positive and false positive rate in the y and x-axis respectively. The true positive rate is the proportion where the model correctly predicts the positive class. Whereas the false positive rate is the proportion in which the model incorrectly predicts the positive class.

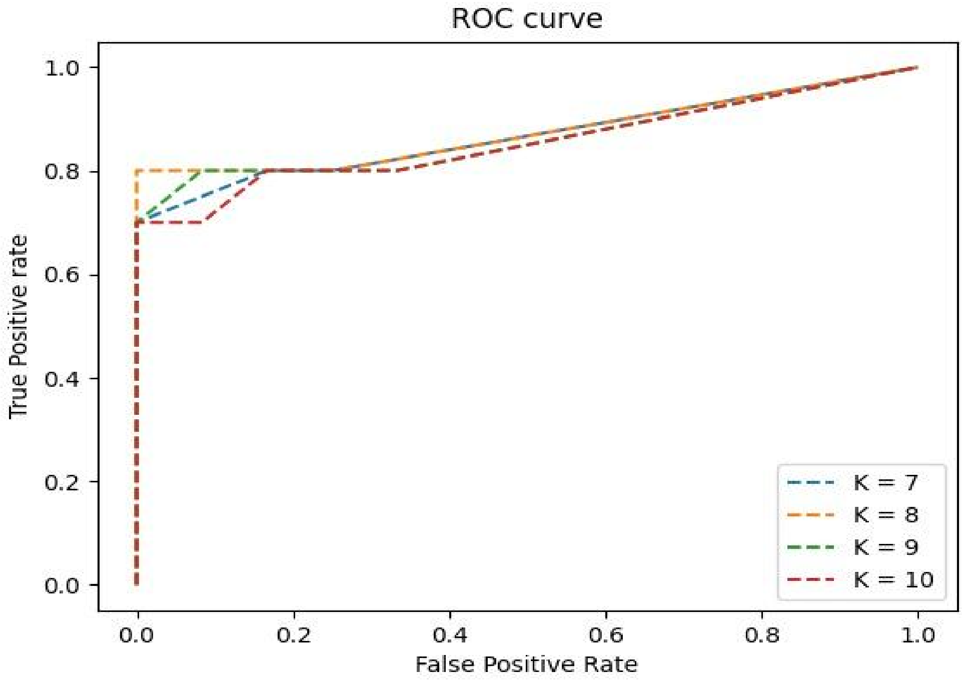


**Figure 3.14: ROC curve for k= 7 → 10.** This curve shows the classification performance of the model for different values of k, from 7 to 10.

As can be seen by the ROC curve above k=8 (orange) (Figure 3.14) performs better than all other values of the independent variable. The next graph (Figure 3.15) isolates this configuration for better visibility.

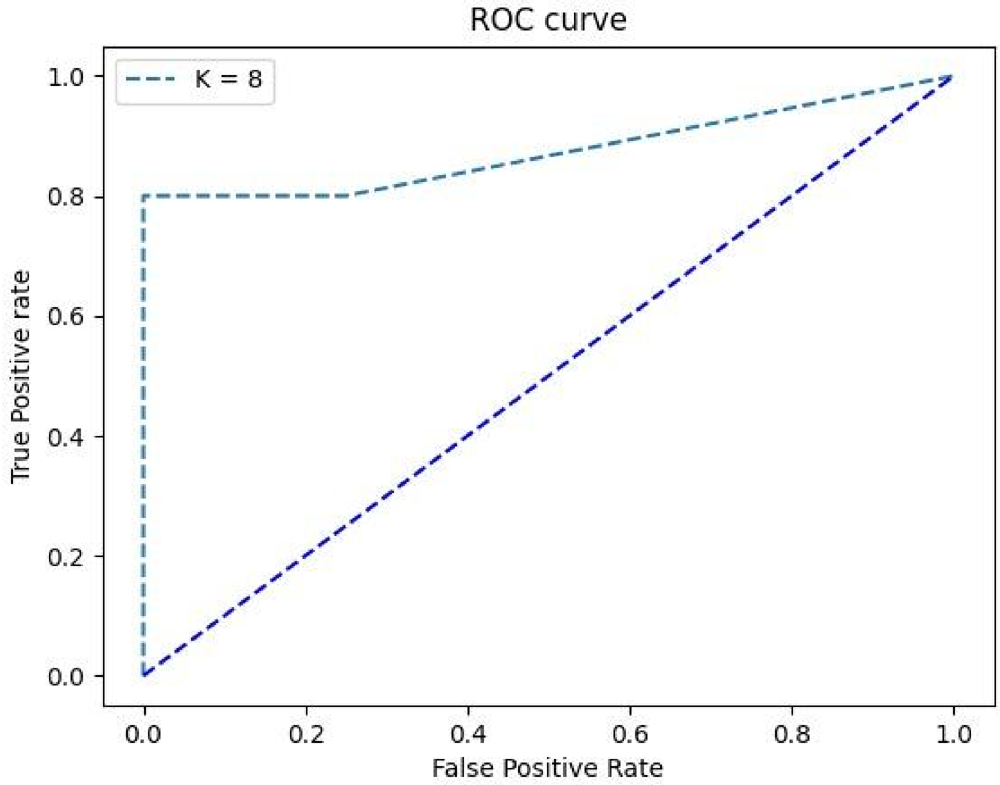


**Figure 3.15: ROC curve for k=8.** This graph clearly shows the classification performance of the model when the value of k is set to 8.

An ROC value of 0.80 is achieved at false positive rate of 0.0 and sustains at this value till false positive rate of 0.30. After this the true positive rate increases linearly with false positive rate till both achieve the final value 1.0 together. This yields an area under the ROC curve of 0.87 which indicates a good performance of the model.

# DISCUSSION

# 4. DISCUSSION

Machine learning is an important area of research in today's era. Different machine learning models are used for different fields like disease diagnosis, image processing, face recognition, text recognition, natural language processing and audio generation etc. Machine learning includes several steps which includes data collection, preparation or preprocessing of data, choosing the right model according to the nature of data and the problem at hand, training the model, evaluating the model by finding the accuracy of the model, parameter tuning and finally making predictions based on the results. Machine learning helps to solve problems that require too much manual effort to complete by speeding up the process. This research tried to find patterns in histopathology images of colon cancer and tried to classify them based on the extracted features from the images. A histopathological image dataset of colon cancer from the Kaggle database is used. These images were H&E stained images in which the nucleus region is bright blue in color whereas the extracellular matrix is in pink color. The stained images help pathologists to identify the region of interest which can be difficult to see with the naked eye. The images are of size 768x768 pixels in size and in JPEG format. The JPEG format is the compressed version of the images to conserve some space in the computer at the expense of image quality. But as the field related to cancer detection and classification is very important, there is a need to have high resolution lossless images like .tiff and .png files for analysis and classification of histopathological images. An image dataset consists of digitized images that have been carefully selected for use in training, testing, and assessing the performance of computer vision and machine learning algorithms. In this research only the nucleus region is focused as cancer cell's nucleus are frequently aberrant in size and form. A cancer cell's nucleus typically differs substantially in size from a normal cell's nucleus and is bigger and darker. This research is divided into eight major steps. First, images are taken from the Kaggle database and then they are preprocessed. During the preprocessing, image normalization technique was used to increase the brightness and contrast. In this research only the nuclear area is the region of interest as it is the most affected area in cancer and a lot of visible differences can be seen easily. Next the region of interest is extracted using color segmentation technique. For color segmentation the algorithm is to match the RGB values of given pixels and select those that fall within a predefined range of two colors, one being on the lighter end of the spectrum and the other on the darker end of the spectrum for colors that correspond to the nucleic region. During color segmentation only the bright blue color is extracted through using a defined range of values as in H&E stained images the nucleus area is in bright blue color. Once these pixels are selected we apply Sobel filter blurring effect to enlarge the area selected by the initial pixel color filtering selection described earlier. This helps increase the extracted area to include small immediate neighboring pixels that might be overlooked due to sharp contrast or noise in the image. The selected area then remains as it is whereas the other parts in the image are turned white. This helped to properly visualize the nuclear regions and to extract the features. This study attempted edge extraction using Canny edge detection technique of OpenCV but as the nuclear regions are enclosed in shape so it had an adverse effect on the output, the algorithm picked up neighboring nuclei as a single region and did not help to properly make boundaries around them. For this purpose the contours are formed around the nucleus which are then converted to the bounding boxes. These bounding boxes are then rotated upright for the proper calculations of the features. These nucleus contours only help to find boundaries without individually calculating the values of every nucleus present in a histopathological image. For this purpose the Stardist algorithm was used to find the instance segmentation of every nucleus in the images. After finding the instance segmentation of the nucleus the features of the nucleus area are calculated. Feature vector contains nucleus mean area, nucleus area standard deviation, nucleus mean height, nucleus mean width, aspect ratio and malignancy. These feature vector values are calculated statistically and then added to the .csv file. After a library is created containing the shape values of the nucleus the graphs are generated which helps to find out the comparison of benign and malignant image feature values. The nucleus mean area of the benign nucleus was seen to be quite less than that of the nucleus mean area of the malignant nucleus. This is because as the cancer starts to develop in the body the nucleus starts deforming. The nucleus mean height of the malignant images is more than those of the benign ones as the nucleus starts to stretch and becomes more elliptical in shape. The aspect ratio showed that the benign nucleus aspect ratio is less than that of the malignant one. If the value is 1 it means that the nucleus in the image is circular but in this case the nucleus were not in a circle but more deformed and elongated in case of the malignant images than those of the benign ones. After calculating the feature vector values the correlation between the feature vector and the malignancy was carried out. Through finding the correlation it was concluded that the nucleus mean area, nucleus aspect ratio and nucleus mean height have more correlation values than that of other feature vectors like the nucleus mean width and nucleus area standard deviation. This feature vector is now reduced in size containing only the useful features while dropping the others not very useful in the classification phase.

The K-NN algorithm is used for the classification purpose. In order to find the accuracy of the model different values of k were used and it was observed that the model was performing best at k=8 with the highest peak in the graph. In this work, input data was divided into two sets. One is termed as training, and it trains the 80% of the input data based on the feature vector containing nucleus mean area, nucleus mean height and aspect ratio. Remaining 20% is termed as testing data, which is labeled and the machine learning model annotates it on the basis of learned feature vector values. Once the model is trained and validated through finding the accuracy it shows that the model is 90.91% accurate.

This study suggests to produce more data related to histopathology images of cancer so more improved models would be readily available for the further analysis and detection of cancers at a much faster rate and with greater accuracy. Additionally, more histological pictures of colon cancer can be added to this model, and additional analysis can be done based on features other than the nucleus that an image may have. This histological imaging data of numerous different cancers must be made available to the public in order for more research in this area of cancer to be pursued.

# CONCLUSION

# 5. CONCLUSION

Given that cancer is the biggest cause of mortality in the world, it is important to focus more on diagnosis and better control of this disease. Although many scientists are working on it, more research is needed in this field to diagnose the condition more quickly and affordably due to restricted resources. The newest technology for disease diagnosis is machine learning. In this study, colon cancer histopathology H&E stained images are utilized to look for patterns that could indicate whether a tumor is benign or malignant. The photos were processed using a variety of machine learning techniques, and then nucleus regions were separated using color segmentation. Stardist algorithm was used to separate the nucleus instances in an image by determining their statistical values, which resulted in a feature vector. The images were then classified as benign or malignant using a K-Nearest Neighbor approach, and the model's accuracy was determined at 90.91%. This colon cancer dataset is in JPEG format and comes from the Kaggle database. Since the JPEG format is compressed, some information is lost during compression that cannot be recovered, potentially lowering the quality of the images. Therefore, it is necessary to create high quality histopathology photographs, which are often in .tiff format. This will make it easier for the scientists to perform analysis tasks with high precision because cancer detection requires greater accuracy in order to save lives.

The traditional method for scanning the histopathology image is performed manually by pathologists who look for patterns in the whole slides. This method is quite time taking and laborious as it takes days and even months for proper diagnosis. Moreover there is a chance of human error in this approach. This developed method can be a great source for the pathologists to find only those images that contain cancerous regions as benign or malignant. This not only saves time but also reduces effort. This software can be used for the early screening purposes as it is 90.91% accurate which will not only help in diagnosis at a faster rate but also can be made as a standard technique used in cancer diagnosis because of its accuracy.

# REFERENCES

# 6. REFERENCES

Chiasserini, D., Biscetti, L., Farotti, L., Eusebi, P., Salvadori, N., Lisetti, V., Baschieri, F., Chipi, E., Frattini, G., Stoops, E., Vanderstichele, H., Calabresi, P., & Parnetti, L. (2016). Performance Evaluation of an Automated ELISA System for Alzheimer's Disease Detection in Clinical Routine. Journal of Alzheimer's Disease, 54(1), 55–67. https://doi.org/10.3233/jad-160298

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: an overview. International Journal of Cancer, 149(4). https://doi.org/10.1002/ijc.33588

Ganesh, V. R., & Ramesh, H. (2017). Effectiveness of contrast limited adaptive histogram equalization technique on multispectral satellite imagery. Idr.nitk.ac.in. https://idr.nitk.ac.in/jspui/handle/123456789/7906

He, L., Long, L. R., Antani, S., & Thoma, G. R. (2012). Histology image analysis for carcinoma detection and grading. Computer Methods and Programs in Biomedicine, 107(3), 538–556. https://doi.org/10.1016/j.cmpb.2011.12.007

Huang, J., Li, J., Li, Z., Zhu, Z., Shen, C., Qi, G., & Yu, G. (2022). Detection of Diseases Using Machine Learning Image Recognition Technology in Artificial Intelligence. Computational Intelligence and Neuroscience, 2022, 1–14. https://doi.org/10.1155/2022/5658641

Hummel, R. (1977). Image enhancement by histogram transformation. Computer Graphics and Image Processing, 6(2), 184–195. https://doi.org/10.1016/s0146-664x(77)80011-7

Kainz, P., Pfeiffer, M., & Urschler, M. (2017). Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. PeerJ, 5, e3874. https://doi.org/10.7717/peerj.3874

Ketcham, D. J., Lowe, R. W., & Weber, J. W. (1974). Image Enhancement Techniques for Cockpit Displays. https://doi.org/10.21236/ada014928

Komura, D., & Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. Computational and Structural Biotechnology Journal, 16, 34–42. https://doi.org/10.1016/j.csbj.2018.01.001

Kumar, A., Singh, S. K., Saxena, S., Lakshmanan, K., Sangaiah, A. K., Chauhan, H., Shrivastava, S., & Singh, R. K. (2020). Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. Information Sciences, 508, 405–421. https://doi.org/10.1016/j.ins.2019.08.072

Long, W., Xia, L., & Wang, X. (2016). A rapid automatic analyzer and its methodology for effective bentonite content based on image recognition technology. China Foundry, 13(5), 322–326. https://doi.org/10.1007/s41230-016-5119-6

Mahesh, B. (2019) Machine Learning Algorithms - A Review. International Journal of Science and Research, 9, 381-386. http://dx.doi.org/10.21275/ART20203995

Mahmood, H., Shaban, M., Indave, B. I., Santos-Silva, A. R., Rajpoot, N., & Khurram, S. A. (2020). Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. Oral Oncology, 110, 104885. https://doi.org/10.1016/j.oraloncology.2020.104885

Nguyen, M., Mikita, G., & Hoda, R. S. (2015). “Intercellular bridges” in a case of well differentiated squamous carcinoma. Diagnostic Cytopathology, 44(2), 121–123. https://doi.org/10.1002/dc.23406

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing, 39(3), 355–368. https://doi.org/10.1016/S0734-189X(87)80186-X

Qaiser, T., Tsang, Y.-W., Taniyama, D., Sakamoto, N., Nakane, K., Epstein, D., & Rajpoot, N. (2019). Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Medical Image Analysis, 55, 1–14. https://doi.org/10.1016/j.media.2019.03.014

Rivera, C., & Venegas, B. (2014). Histological and molecular aspects of oral squamous cell carcinoma (Review). Oncology Letters, 8(1), 7–11. https://doi.org/10.3892/ol.2014.2103

Sund, T., & Møystad, A. (2006). Sliding window adaptive histogram equalization of intraoral radiographs: effect on image quality. Dentomaxillofacial Radiology, 35(3), 133–138. https://doi.org/10.1259/dmfr/21936923

Taylor, P. R., Abnet, C. C., & Dawsey, S. M. (2013). Squamous Dysplasia--The Precursor Lesion for Esophageal Squamous Cell Carcinoma. Cancer Epidemiology Biomarkers & Prevention, 22(4), 540–552. https://doi.org/10.1158/1055-9965.epi-12-1347

Umbaugh, S. E. (2011). Digital image processing and analysis : human and computer vision applications with CVIPtools. Crc Press.

Webb, S. (2018). Deep learning for biology. Nature, 554(7693), 555–557. https://doi.org/10.1038/d41586-018-02174-z

Weigert, M., Schmidt, U., Haase, R., Sugawara, K., & Myers, G. (2020). Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/wacv45572.2020.9093435

Xin, B., Zhang, J., Zhang, R., & Wu, X. (2016). Color texture classification of yarn-dyed woven fabric based on dual-side scanning and co-occurrence matrix. Textile Research Journal, 87(15), 1883–1895. https://doi.org/10.1177/0040517516660886

Zhang, X., Xing, F., Su, H., Yang, L., & Zhang, S. (2015). High-throughput histopathological image analysis via robust cell segmentation and hashing.

Medical Image Analysis, 26(1), 306–315. https://doi.org/10.1016/j.media.2015.10.005