**Abstract**

Cancer is a group of diseases involving abnormal cell growth

According to American cancer society Cancer continues to be the second most common cause of death in the US, after heart disease. A total of 1.9 million new cancer cases and 609,360 deaths from cancer are expected to occur in the US in 2022, which is about 1,670 deaths a day.

no permanent cure has been developed to combat cancer, early detection is crucial for treatment and survival of patient.

Image recognition and deep learning have been used effectively in detection and treatment of several dangerous diseases, helping in early diagnosis and treatment.

The risk of death from cancer dropped by about 2% a year from 2015 through 2019 compared to 1% a year during the 1990s. Accelerating declines in the cancer death rate show the power of prevention, screening, early diagnosis, treatment.

Deep learning can be used to analyze features allowing detection of breast cancer.

Two of the most common imaging used in breast cancer detection are histopathology and mammography.

\*In our research we aim to compare different deep learning methods on each type of the two imaging, while analyzing the results of the different methods for detection and classification

Key Words: Breast Cancer; Image recognition; Deep Learning; classification

**1. Introduction**

Breast cancer is the second most diagnosed cancer worldwide.[1]

\*(ask ronen if cite is needed or reference is enough)

. Breast cancer occurs in four main types: normal, benign, in-situ carcinoma and invasive carcinoma [2].

In situ carcinoma, the cancer does not effect other organs other than mammary duct lobule system. Benign is not classified as a harmful cancer and involves a minor change in the breast structure. Invasive carcinoma is the deadliest type out of all the four main breast cancer types cause it can spread out to all other organs.

Breast cancer can be diagnosed using one of two approaches: histopathological image analysis or mammography.

Histopathological images are microscopic images of breast tissue that are extremely useful in early treatment of the cancer.

Mammography is specialized medical imaging that uses a low-dose x-ray system to see inside the breasts. A mammography exam, called a mammogram, aids in the early detection and diagnosis of breast diseases in women.

The main difference between:  
mammography is an earlier type of imaging, before breast tissue is collected for histopathology, an x ray image inside the breasts allows us to search for lumps indicating cancer cells, if there is an indication for breast cancer, breast tissue is collected for analysis under microscope for more accurate diagnosis.

We aim to compare and find which type of deep learning methods will satisfy with most accuracy for each type while analyzing the results.

CNN-  
https://www.analyticsvidhya.com/blog/2021/06/breast-cancer-classification-using-deep-learning/

RCNN-  
https://github.com/riblidezso/frcnn\_cad

**SVM**-  
https://towardsdatascience.com/case-study-breast-cancer-classification-svm-2b67d668bbb7

**KNN**-  
<https://www.kaggle.com/code/nsaravana/breast-cancer-using-knn-algorithm/notebook>  
<https://www.codingninjas.com/codestudio/library/breast-cancer-classification-using-knn>  
https://github.com/Manishnir/Breast-Cancer-Prediction-using-KNN

RNN WITH LSTMS  
<https://www.kaggle.com/code/data13/rnn-model-for-breast-cancer-classification/notebook>

These are the methods which will use for our image recognition and classification. We will use each method on mammography and histopathology dataset

We will study each algorithm and make an assumption based on the algorithm aspects on how he will compare against other on each type of images.

The remaining of this paper is structured as follows: Section 2 presents related work which includes surveys conducted in breast cancer area. Section 3 explains the methodology used to conduct this research. Section 4 presents the obtained results and related discussions. Lastly, Section 5 concludes the paper and suggests future research directions.

References:

1 S. Germano and L. O'Driscoll, Curr. Cancer Drug Targets, 2009, 9, 398–418.

**2 Background and Related Work**

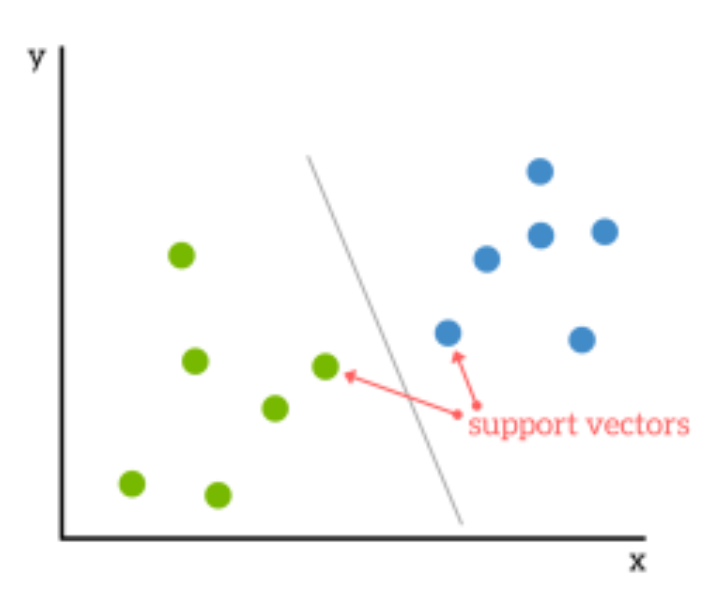
**2.1 Background – SVM (Support Vector Machine)**

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems and such.

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes.

**2.1.1 Support Vectors**

Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing **hyperplane**. Because of this, they can be considered the critical elements of a data set.



**2.1.2 Hyperplane**

As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

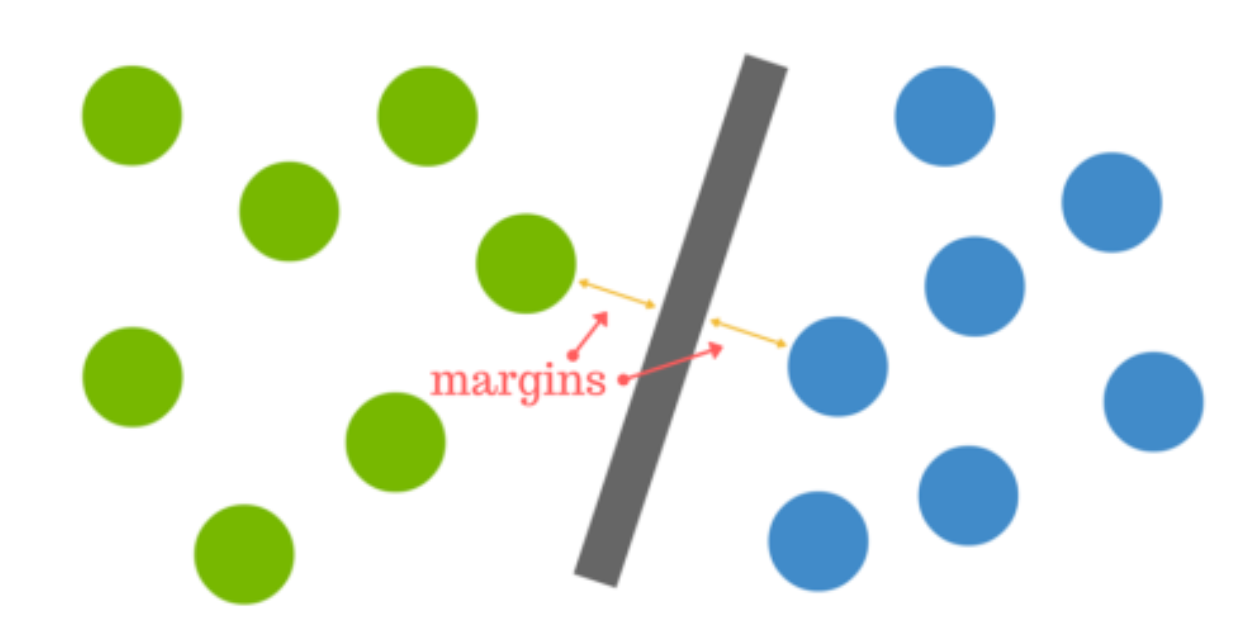
Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

**2.1.2.1 Definition**

**2.1.2.2 Finding the right hyperplane**

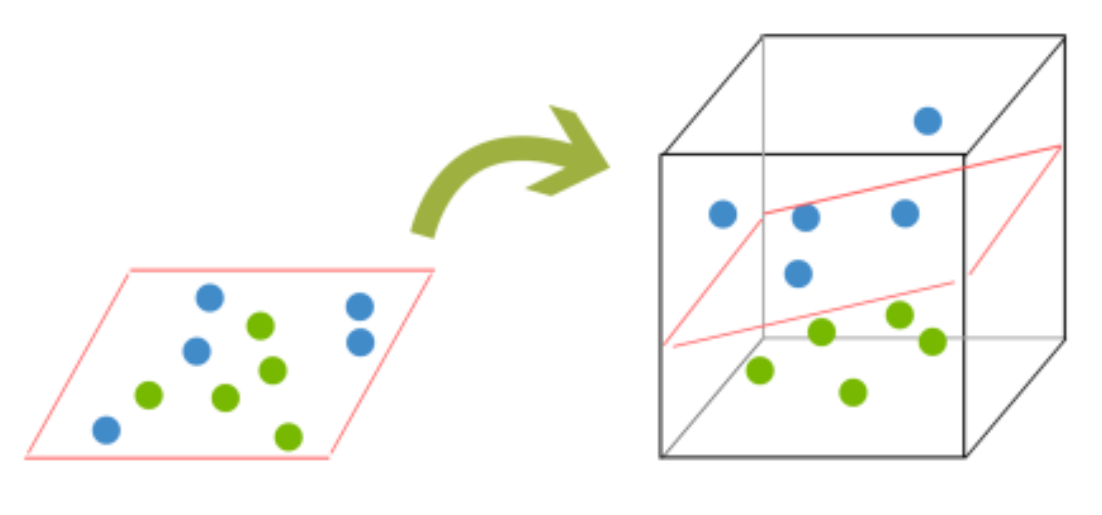
The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible **margin** between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.



**2.1.2.3 Solving whenever there is no clear hyperplane**

This is where it can get tricky. Data is rarely ever as clean as our simple example above. A dataset will often look more like the jumbled balls below which represent a linearly non separable dataset.

In order to classify a dataset like the one above it’s necessary to move away from a 2d view of the data to a 3d view. Explaining this is easiest with another simplified example. Imagine that our two sets of colored balls above are sitting on a sheet and this sheet is lifted suddenly, launching the balls into the air. While the balls are up in the air, you use the sheet to separate them. This ‘lifting’ of the balls represents the mapping of data into a higher dimension. This is known as **kernelling**.



Because we are now in three dimensions, our hyperplane can no longer be a line. It must now be a plane as shown in the example above. The idea is that the data will continue to be mapped into higher and higher dimensions until a hyperplane can be formed to segregate it.

**2.1.3 Pros & Cons of SVM**

Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it’s common to have access to a dataset of at most a couple of thousands of tagged samples. SVM have its own disadvantage as well: Isn't suited to larger datasets as the training time with SVMs might be high, less effective on noisier datasets with overlapping classes.

**2.1.4 Uses of SVM**

SVM is used for text classification tasks such as category assignment, detecting spam and sentiment analysis. It is also commonly used for image recognition challenges, performing particularly well in aspect-based recognition and color-based classification. SVM also plays a vital role in many areas of handwritten digit recognition, such as postal automation services.

**2.2 Background – KNN (K Nearest Neighbors)**

K-nearest neighbors (KNN) is a type of supervised learning machine learning algorithm and is used for both regression and classification task.

KNN is used to make predictions on the test data set based on the characteristics of the current training data points. This is done by calculating the distance between the test data and training data, assuming that similar things exist within close proximity.

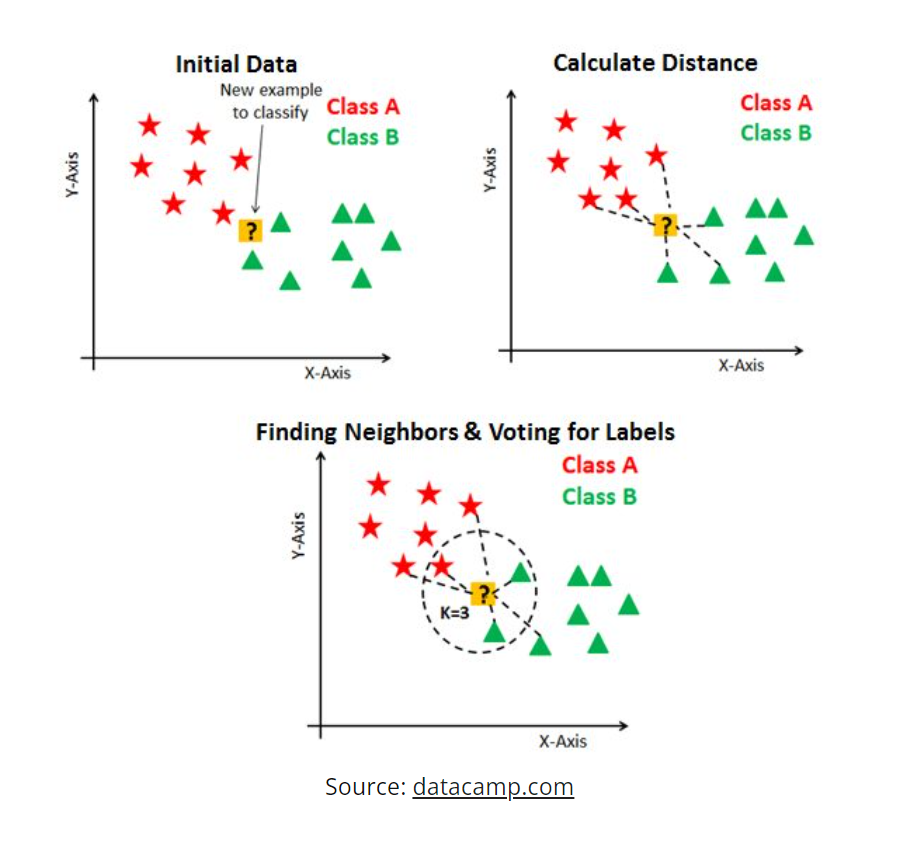
The algorithm will have stored learned data, making it more effective at predicting and categorizing new data points. When a new data point is inputted, the KNN algorithm will learn its characteristics/features. It will then place the new data point at closer proximity to the current training data points that share the same characteristics or features.

**2.2.1 What is K for**

The ‘K’ in KNN is a parameter that refers to the number of nearest neighbors. K is a positive integer and is typically small in value and is recommended to be an odd number.   
In Layman's terms, the K-value creates an environment for the data points. This makes it easier to assign which data point belongs to which category.

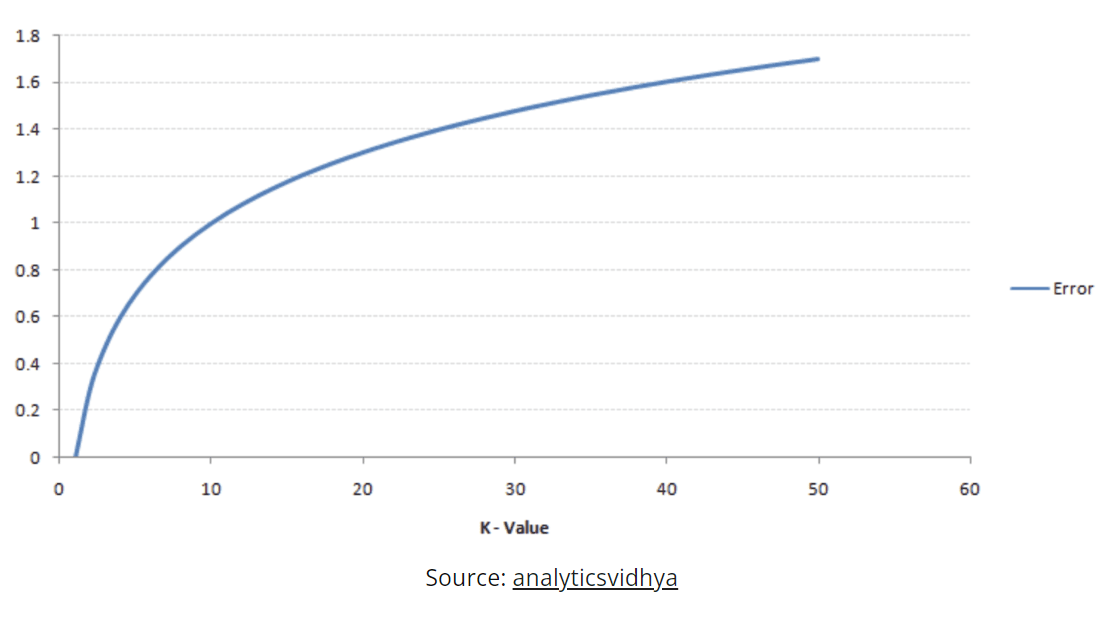
The example below shows 3 graphs. The first, the ‘Initial Data’ is a graph where data points are plotted and clustered into classes, and a new example to classify is present. In the ‘**Calculate Distance**’ graph, the distance from the new example data point to the closest trained data points is calculated. However, this still does not categorize the new example data point. Therefore, using k-value, essentially created a neighborhood where we can classify the new example data point.

We would say that k=3 and the new data point will belong to Class B as there are more trained Class B data points with similar characteristics to the new data point in comparison to Class A.



The k-value is typically a small number because, as we increase the k-value, the error rate also increases.

The below graph shows this:



However, If the k-value is small then it causes a low bias but a high variance, leading to overfitting of the model.

**2.2.1.1 Choosing the K**

To get the right K, you should run the KNN algorithm several times with different values of K and select the one that has the least number of errors.

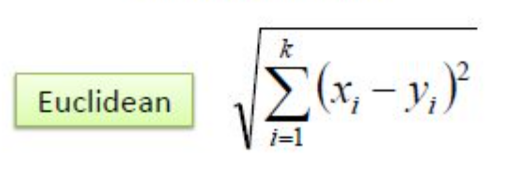
The right K must be able to predict data that it hasn’t seen before accurately.

**2.2.2 Calculate the distance**

Most common methods used to calculate this distance in KNN are **Euclidian**, **Manhattan** and **Minkowski**.

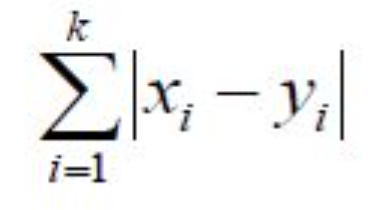
**2.2.2.1 Euclidian Distance**

Euclidean Distance is the distance between two points using the length of a line between the two points. The formula for Euclidean Distance is the square root of the sum of the squared differences between a new data point (x) and an existing trained data point (y).



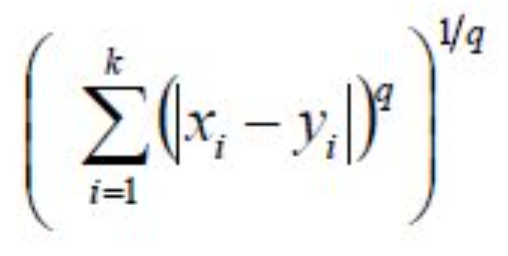
**2.2.2.2 Manhattan Distance**

Manhattan Distance is the distance between two points is the sum of the absolute difference of their Cartesian coordinates. The formula for Manhattan Distance is the sum of the lengths between a new data point (x) and an existing trained data point (y) using a line segment on the coordinate axes.

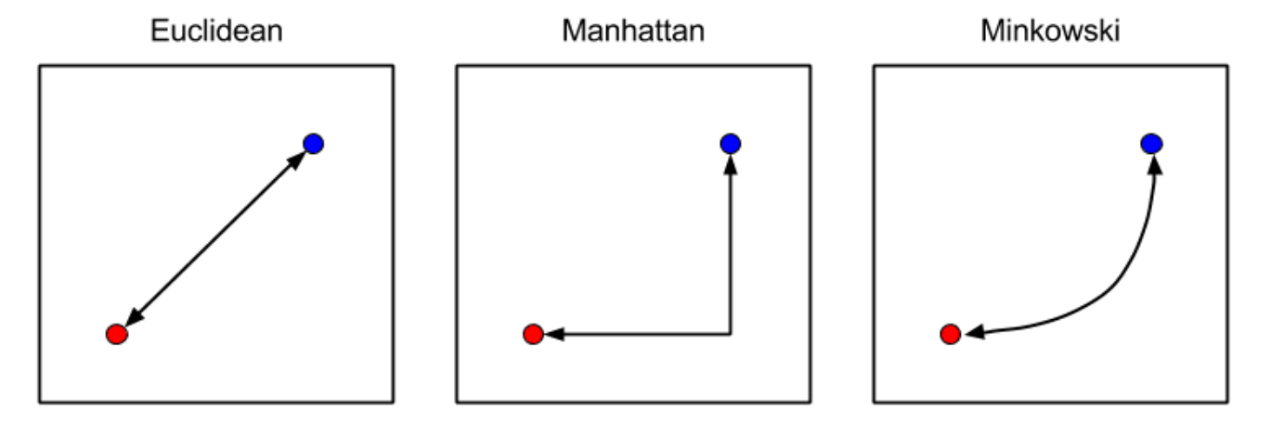


**2.2.2.3 Minkowski Distance**

Minkowski Distance is the distance between two points in the normed vector space and is a generalization of the Euclidean distance and the Manhattan distance. In the formula for ​​Minkowski Distance when p=2, we get Euclidian distance, also known as L2 Distance. When p=1 we get Manhattan distance, also known as L1 distance, city-block distance, and LASSO.



The image below explains the difference between the three:



**2.2.3 Pros & Cons of KNN**

The main advantages for the KNN algorithm: No assumptions about the data, Versatile - useful for classification or regression, Simplicity – easy to explain and understand the concept.

Some of the disadvantages are: High memory requirement, Prediction stage can be slow, Stores most of the training data.

**2.2.4 Uses of SVM**

It’s used in many different areas, such as handwriting detection, image recognition, and video recognition. KNN is most useful when labeled data is too expensive or impossible to obtain, and it can achieve high accuracy in a wide variety of prediction-type problems.