***Importance of machine learning in the fields of cancer detection and diagnosis***

Solanki Acharya

Contents

[Section 1: Introduction 2](#_Toc74091085)

[Section 2: Basics of ML and its types 2](#_Toc74091086)

[Section 3: Importance of ML in Medical Imaging for Cancer Diagnosis 7](#_Toc74091087)

[Section 4: ML techniques in cancer diagnostics 9](#_Toc74091088)

[Section 5: Drawbacks/area of improvement in ML 22](#_Toc74091089)

[Section 6: Conclusion 27](#_Toc74091090)

[References 28](#_Toc74091091)

# Section 1: Introduction

Cancer is one of the most life threatening diseases with high morbidity and mortality rate and has a major impact on society across the world. Cancer statistics describe what happens in large groups of people and provide a picture in time of the burden of cancer on society.

Previously many of statistical studies examined the things such as number of people are diagnosed and die from cancer each year, the number of people who are currently living after a cancer diagnosis, the average age at diagnosis, and the numbers of people who are still alive at a given time after diagnosis. It also informs about differences among groups defined by age, sex, racial/ethnic group, geographic location, and other categories.

The projected number of patients with cancer in India is 1,392,179 for the year 2020, and the common 5 leading sites are breast, lung, mouth, cervix uteri, and tongue. Trends in cancer incidence rate showed an increase in all sites of cancer in both sexes and were high in Kamrup urban (annual percent change, 3.8%; *P* < .05). The majority of the patients with cancer were diagnosed at the locally advanced stage for breast (57.0%), cervix uteri (60.0%), head and neck (66.6%), and stomach (50.8%) cancer, whereas in lung cancer, distant metastasis was predominant among males (44.0%) and females (47.6%).

By looking at cancer rates over time, we can track changes in the risk of developing and dying from specific cancers as well as cancer overall.

Because of such high incidence and heavy load in the laboratories across the world, there is an eager need to work on computer based diagnostics in cancer studies applying some machine learning methods. Further automated cancer diagnosis works on quantitative methods. Innovation in digital imaging using machine learning and computational utility in diagnosis can assist pathologists to increase diagnostic precision and inter observer reliability and lower the current workload.

Machine learning is all about automatically building new algorithms or use existing ones to learn from data, in order to build generalizable and highly accurate predictive or classifier models that give accurate predictions, or to find patterns, particularly with new and unseen similar data.

So, as a whole, Machine learning is a tool that can be used to enhance humans’ abilities to solve problems and make informed inferences on a wide range of problems, from helping diagnose diseases to coming up with solutions for global climate change.

# Section 2: Basics of ML and its types

***Basics of Classical Machine Learning***

Machine learning is a branch of artificial intelligence which includes mathematical, statistical, probabilistic and optimization techniques that allows computers to learn from past examples and to detect hard to discern patterns from large, noisy or complex data sets. This capability is particularly well suited to biomedical applications like cancer diagnosis and detection. Machines that learn are useful to humans because, with all of their processing power, they’re able to more quickly highlight or find patterns in big (or other) data that would have otherwise been missed by human beings.

The widely accepted formal definition of machine learning as stated by field pioneer Tom M. Mitchell is

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E .”

* 1. **Classical ML techniques**

There are 4 types of Classical Machine Learning Methods. They are:

* ***Supervised Learning:*** Basically, this is learning the mapping from input set to output set. Supervised learning is the concept of function approximation, where basically we train an algorithm and in the end of the process we pick the function that best describes the input data, the one that for a given X makes the best estimation of y (X -> y). Supervised learning algorithms try to model relationships and dependencies between the targetprediction output and the input features such that we can predict the output values for new data based on those relationships which it learned from the previous data sets.

There are two types of supervised learning:

1. *Classification (Categorical Output):* This is the process where incoming data is labelled based on past data samples and manually trains the algorithm to recognize certain types of objects and categorize them accordingly. The system or machine has to know how to differentiate types of information, perform an optical character, image, or binary recognition (whether a particular bit of data is compliant or non-compliant to specific requirements in a manner of “yes” or “no”).
2. *Regression (Continuous Output):* This is the process of identifying patterns and calculating the predictions of continuous outcomes. The system has to understand the numbers, their values, grouping (for example, heights and widths), etc.

The most widely used supervised algorithms are:

* Linear Regression
* Logistical Regression
* Random Forest
* Gradient Boosted Trees
* Support Vector Machines (SVM)
* Neural Networks
* Decision Trees
* Naive Bayes
* Nearest Neighbour
* ***Unsupervised Learning:*** This method basically discovers patterns in data. This is the one that does not involve direct control of the developer. If the target of supervised machine learning is that we know the results and need to sort out the data, then in the case of unsupervised machine learning algorithms the desired results are unknown and yet to be defined. Another big difference between the two is that supervised learning uses labelled data exclusively, while unsupervised learning feeds on unlabelled data.

The unsupervised machine learning algorithm is used for exploring the structure of the information, extracting valuable insights, detecting patterns, implementing this into its operation to increase efficiency.

The types of Unsupervised Learning are:

*Clustering (Cohesive Grouping):* It is an exploration of data used to segment it into meaningful groups (i.e., clusters) based on their internal patterns without prior knowledge of group credentials. The credentials are defined by the similarity of individual data objects and also aspects of their dissimilarity from the rest (which can also be used to detect anomalies).

*Association:* Frequent co-occurrence i.e. analysing how likely two events can occur simultaneously.

*Dimensionality reduction:*  If there is a lot of noise in the incoming data, Machine learning algorithms use dimensionality reduction to remove this noise while distilling the relevant information.

The most widely used algorithms are:

* k-means clustering
* t-SNE (t-Distributed Stochastic Neighbour Embedding)
* PCA (Principal Component Analysis)
* Association rule
* ***Semi-supervised Learning:*** Semi-supervised learning algorithms represent a middle ground between supervised and unsupervised algorithms. In essence, the semi-supervised model combines some aspects of both into a thing of its own.

Here’s how semi-supervised algorithms work: A semi-supervised machine-learning algorithm uses a limited set of labelled sample data to shape the requirements of the operation (i.e., train itself). The limitation results in a partially trained model that later gets the task to label the unlabelled data. Due to the limitations of the sample data set, the results are considered pseudo-labelled data. Finally, labelled and pseudo-labelled data sets are combined, which creates a distinct algorithm that combines descriptive and predictive aspects of supervised and unsupervised learning.

Semi-supervised learning uses the classification process to identify data assets and the clustering process to group it into distinct parts.

* ***Reinforcement Learning:*** Thislearning is all about developing a self-sustained system that, throughout contiguous sequences of tries and fails, improves itself based on the combination of labelled data and interactions with the incoming data. Reinforced ML uses the technique called exploration/exploitation. The mechanics are simple - the action takes place, the consequences are observed, and the next action considers the results of the first action. In the centre of reinforcement learning algorithms are reward signals that occur upon performing specific tasks. In a way, reward signals are serving as a navigation tool for the reinforcement algorithms. They give it an understanding of right and wrong course of action.

Most common reinforcement learning algorithms include:

* Q-Learning
* Temporal Difference (TD)
* Monte-Carlo Tree Search (MCTS)
* Asynchronous Actor-Critic Agents (A3C)

***Basics of Deep Learning***

Deep learning is a subfield of machine learning that structures algorithms in layers to create an "artificial neural network” that can learn and make intelligent decisions on its own.

A [neural network](https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-neural-network) is structured like the human brain and consists of artificial neurons, also known as nodes. These nodes are stacked next to each other in three layers:

* The input layer
* The hidden layer(s)
* The output layer

In practical terms, deep learning is just a subset of machine learning. In fact, deep learning is machine learning and functions in a similar way. However, its capabilities are different.

While basic machine learning models do become progressively better at whatever their function is, they still need some guidance. If an AI algorithm returns an inaccurate prediction, then an engineer has to step in and make adjustments. With a deep learning model, an algorithm can determine on its own if a prediction is accurate or not through its own neural network.

A deep learning model is designed to continually analyze data with a logic structure similar to how a human would draw conclusions. To achieve this, deep learning applications use a layered structure of algorithms called an **artificial neural network**. The design of an artificial neural network is inspired by the biological neural network of the human brain, leading to a process of learning that’s far more capable than that of standard machine learning models.

* 1. **DL techniques**

Deep learning models make use of several algorithms. While no one network is considered perfect, some algorithms are better suited to perform specific tasks.

Most commonly known DL techniques are as follows:-

1. Convolutional Neural Networks (CNNs)
2. Long Short Term Memory Networks (LSTMs)
3. Recurrent Neural Networks (RNNs)
4. Generative Adversarial Networks (GANs)
5. Radial Basis Function Networks (RBFNs)
6. Multilayer Perceptrons (MLPs)
7. Self-Organizing Maps (SOMs)
8. Deep Belief Networks (DBNs)
9. Restricted Boltzmann Machines( RBMs)
10. Autoencoders

# Section 3: Importance of ML in Medical Imaging for Cancer Diagnosis

* **Analysis of Microscopy Data**

Machine Learning (ML) tries to implicitly learn information from available data. The computer learns from example data, which describe the problem as generally as possible, in connection with an associated target quantity such as an image or material property. The aim of ML methods is to recognize patterns in data and to derive underlying principles from them. These patterns can also represent complex, non-linear relationships. Machine Learning is particularly suitable for problems in which cause-and-effect relationships are not completely known or can hardly be grasped by humans due to too many influencing variables. For ML to work, it needs a sufficient amount of high-quality data, as well as experts to label - that is, to annotate the data to train the algorithms.

* **Image classification**

Image classification is one of the most widely used functions of deep learning. Deep learning has been applied to recognize and count particular biological objects from phenotype images such as cells, microarray, and bacterial colonies. Experiments with varying size of training and test data show a 90% average success rate.

* **Image Segmentation & Detection of Region of Interest**

With the development and progress of faster algorithms and stronger hardware support, deep learning has further developed and is able to make predictions for every single pixel. Digital pathological image analysis using computer assisted diagnosis can track tumour region, scoring of immune staining, cancer staging, mitosis detection, gland segmentation, and detection and quantification of vascular invasion.

* **Image enhancement**

In the field of computer vision, deep learning technique has been used for improving the image quality of photographic images, such as image super-resolution (i.e., generating high-resolution image with rich texture details from low-resolution image) and image synthesis. Similar techniques have been quickly applied for biological microscopic image analysis. CNN-based image super-resolution methods were used to improve the image resolution of conventional optical microscopic images, mobile phone microscopic device, photo activated localization microscopy, and stochastic optical reconstruction microscopy. In addition to CNN, generative adversarial network (GAN) is another powerful framework for learning microscopic image patterns and compensating for the missing image information of the low-resolution microscopic images.

* **Discovering New Clinical & Pathological Relationships**

It is important to correlate the morphological features of cancers with their clinical behaviour. For example, tumour grading is important in planning treatment and determining a patient’s prognosis for certain types of cancer. By analyzing the relationship between large amount of digital information such as, digital pathological images, MRI and CT images new pathological relationships, for example, the relationship between the morphological characteristic and the somatic mutation of the cancer, can be found. However, since the amount of data is very large, it is not really possible for pathologists and researchers to analyze all the relationships manually by looking at the specimens. This is where the machine learning technology comes in.

* **Liquid Biopsy and Monitoring the Effects of Chemotherapy using Machine Learning**

A new biomedical technology in medical cancer scanning is a liquid biopsy based on collecting circulating tumour cells and subjecting them to genetic analysis. This way, it is possible to monitor the evolution of cancer during drug treatment. This is a very promising combination of technologies that can not only automate the diagnosis and treatment, but also has potential to help discover new biomarkers correlated with chemotherapy.

* **Therapy Prediction & Personalized Medicine**

Deep learning (DL) is a specialized machine learning approach based on multiple-layered structures algorithms of artificial neurons, which are able to process information and learn by adjusting the weights at each synapse, enabling the performance of an intelligent task with high precision. Furthermore, these deep networks are able to ‘learn’ from large amounts of data and, similar to the human brain, can generalize concepts and apply these to new data with high accuracy. As a result, current DL systems are able to integrate and model heterogeneous data from an individual patient, across modalities and time, allowing better predictions and recommending treatment options to each patient's individual characteristics and needs. Also DL provides a novel approach, which enables developing high‑accuracy, multi‑modal predictive models that allow the implementation of the personalized medicine vision in the near future.

# Section 4: ML techniques in cancer diagnostics

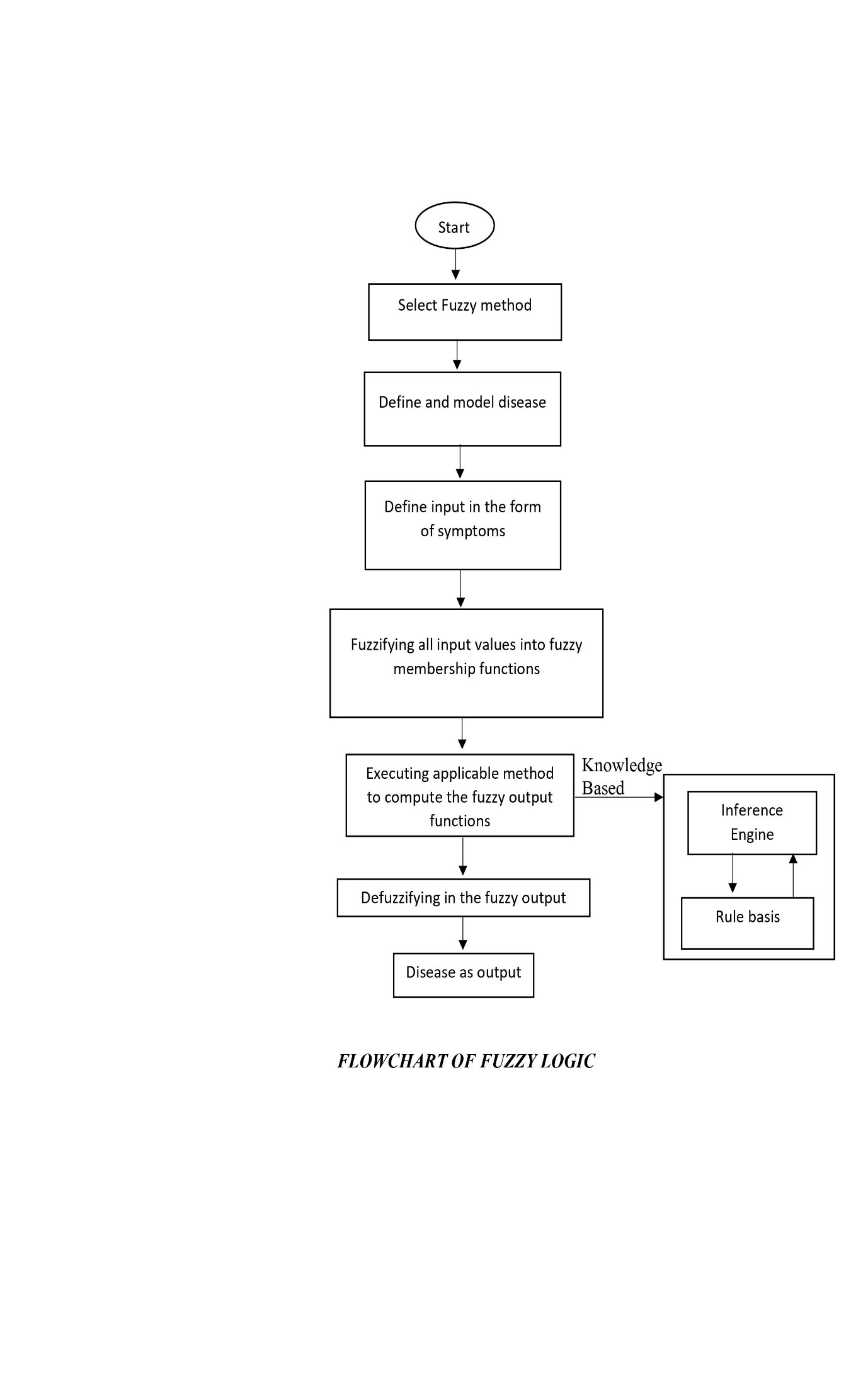
**MOST COMMON AND CONVENTIONAL TECHNIQUES OF ML PREVIOUSLY USED FOR CANCER PREDICTION**

* ***Fuzzy logic (Cancer Type : bladder, breast, throat, lung)***

The fuzzy logic approach gives machines the ability of processing special data of humans and of working by benefiting from their experiences and foresights. While bringing in this ability, it uses symbolical expressions instead of numerical expressions. The transfer of these symbolic expressions to the machines is based on a mathematical basis. This mathematical basis is the fuzzy logic sets theory and fuzzy logic is based on this. The basis of a fuzzy logic controller is this kind of verbal expression and the logical relationships between them.

Fuzzy logic in its simplest terms expands the dichotomy of true or not true to include a range of degrees of truth answers in between. Introducing partial truths, fuzzy logic is more appropriate in medicine where diagnosis implies complex data involving several levels of uncertainty and imprecision. An important advantage that can stand alone in justifying the use of fuzzy logic in medicine is the ability of this machine algorithm to introduce into the process of decision linguistic terms, easier for human users to understand and communicate with. Systems that use fuzzy logic can produce effective results based on indefinite verbal knowledge, like humans. In fuzzy logic, information is verbal phrases, such as ‘big’, ‘small’, ‘very’, or ‘few’, instead of numeric values. If a system’s behaviour can be expressed by rules or requires very complex nonlinear processes, the fuzzy logic approach can be applied.

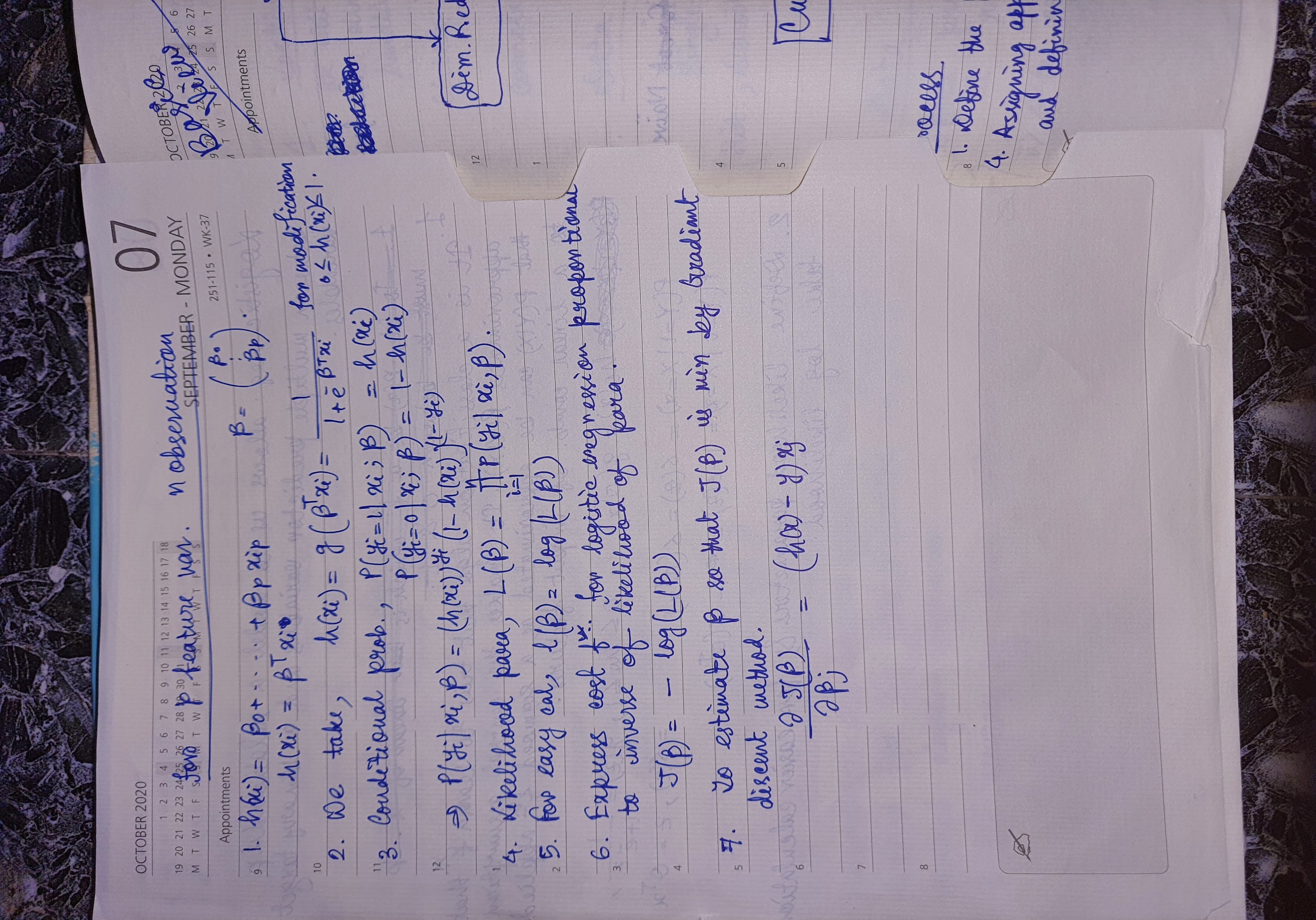
During the application of the fuzzy logic controller, the mathematical modelling of the system is not essential. The transfer of verbal expressions to the computer is based on a mathematical basis. This mathematical basis is named fuzzy sets theory and fuzzy logic. Fuzzy logic expresses multilevel procedures in the [0,1] range, unlike the 2 levels of (0,1) as in classical logic. Fuzzy logic has the ability to conduct procedures according to information that is not fully known or entered incompletely.



* ***Logistic Regression***

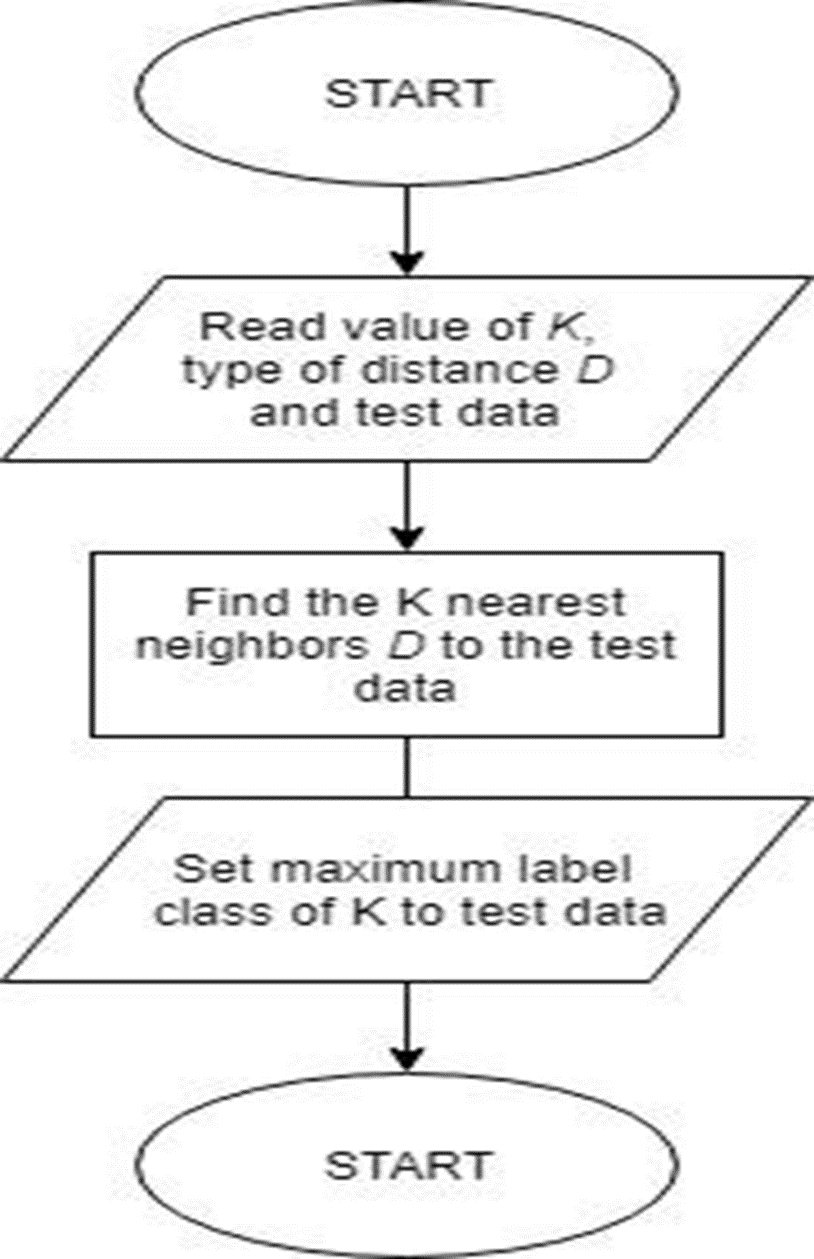
Logistic Regression allows us to model a relationship between multiple predictor variables and a binary target variables.

It is a classification algorithm that tries to learn a function that approximates P(Y/X). It makes the central assumption that P(Y/X) can be approximated as a sigmoid function applied to a linear combination of input features.

******

* ***KNN***

Mostly used for classification of data on the basis how its neighbourhood are classified. K in KNN is a parameter that refers to the number of the nearest neighbours to include in the majority voting process.

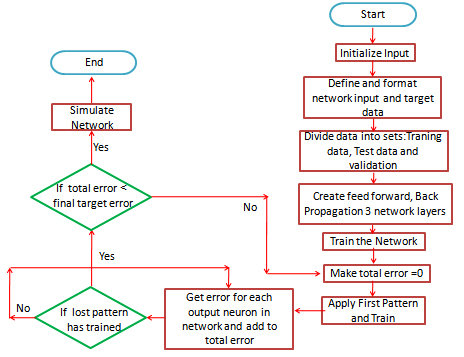


***FLOWCHART OF KNN***

***ANN (Cancer Type : bladder, breast, brain, cervical,*** ***colorectal,*** ***esophageal, liver, lung, lymphoma, head/neck, prostate, skin, stomach, throat, thoracic)***

The artificial neural networks (ANNs) are statistical models where the mathematical structure reproduces the biological organisation of neural cells simulating the learning dynamics of the brain. Although definitions of the term ANN could vary, the term usually refers to a neural network used for non-linear statistical data modelling. Similar to neurons in the brain, ANN also consists of neurons which are arranged in various layers. Feed forward neural network is a popular neural network which consists of an input layer to receive the external data to perform pattern recognition, an output layer which gives the problem solution, and a hidden layer is an intermediate layer which separates the other layers. The adjacent neurons from the input layer to output layer are connected through acyclic arcs. So, basically, the input layer receives a data set (related to the research question) and or more hidden layers (connected in a specified hierarchical manner) synthesise this data. The output layer both receives the hidden layer(s) and generates the answer to the research question. The ANN uses a training algorithm to learn the datasets which modifies the neuron weights depending on the error rate between target and actual output. In general, ANN uses the back propagation algorithm as a training algorithm to learn the datasets.

ANNs have unique properties including robust performance in dealing with noisy or incomplete input patterns, high fault tolerance, and the ability to generalise from the training data which make them capable of clustering, function approximation, forecasting and association, and performing massively parallel multifactorial analyses for modelling complex patterns which leads them to be applied hugely within medical decision-making fields. Neural network techniques are very useful for detection and monitoring of cancer. Artificial neural network had been proven a very effective method for pattern recognition and this made them very useful for diagnosis of cancer disease at very early stages.



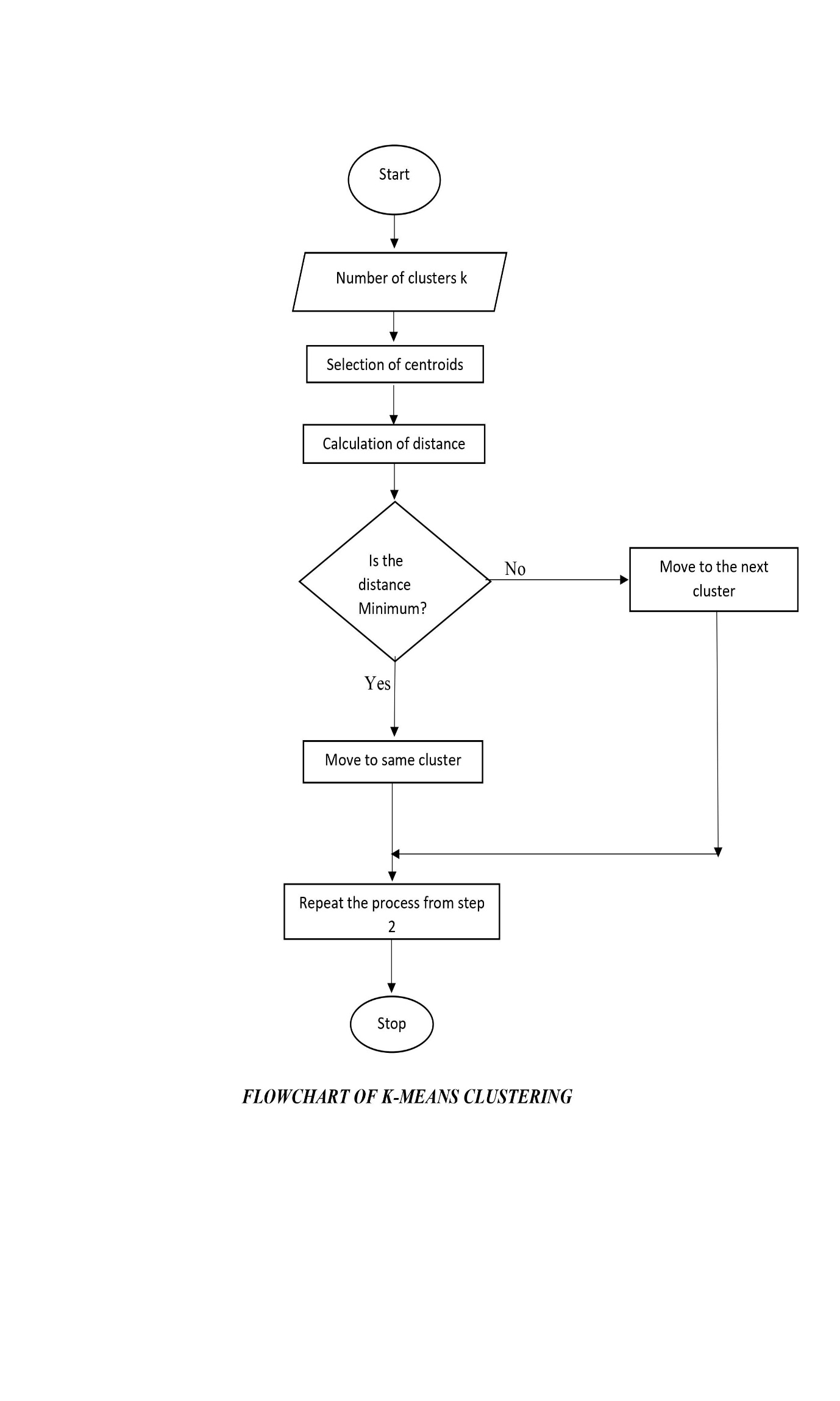
***FLOWCHART OF ANN***

* ***Clustering (Cancer Type : pleural mesothelioma, breast, colorectal, lymphoma)***

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.

Two most common clustering methods used in Automated Cancer Diagnosis are :-

[**K-means clustering algorithm**](https://www.geeksforgeeks.org/k-means-clustering-introduction/) - It is the simplest unsupervised learning algorithm that solves clustering problem. K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



**Principal Component Analysis (PCA)** - PCA is a powerful data reduction method. It converts high-dimensional data into a few numbers of variables, called a set of uncorrelated principal components (PCs), with a hope that this low-dimensional space (PCs) will well represent the original data. Each PC is a linear function of original variables to maximize variance in the original set. The weight in the linear function is called loading coefficient of the PCs and is the basis of the low-dimensional space. Thus, the constructed PCs become new coordinates in the new system. Alternatively, PCs can be viewed as the points that the original observations are projected onto a low-dimensional space with the aim to be close as possible to the original data set. For example, gene expression data include thousands of genes (original data). It is a challenge to analyze this raw high-dimensional data. However, if we can use PCA to reduce data into a few PCs, say 3 PCs, standard statistical methods, such as linear regression or logistic regression models, can be easily applied.

PCA reduces the dimensionality by finding a few orthogonal linear combinations of the original variables with the largest variance. Basically, PCA tries to capture the maximal variance within the data. In particular, it finds a pair of orthogonal vectors that define a lower dimensional space which captures as much as variance as possible from the original dataset.

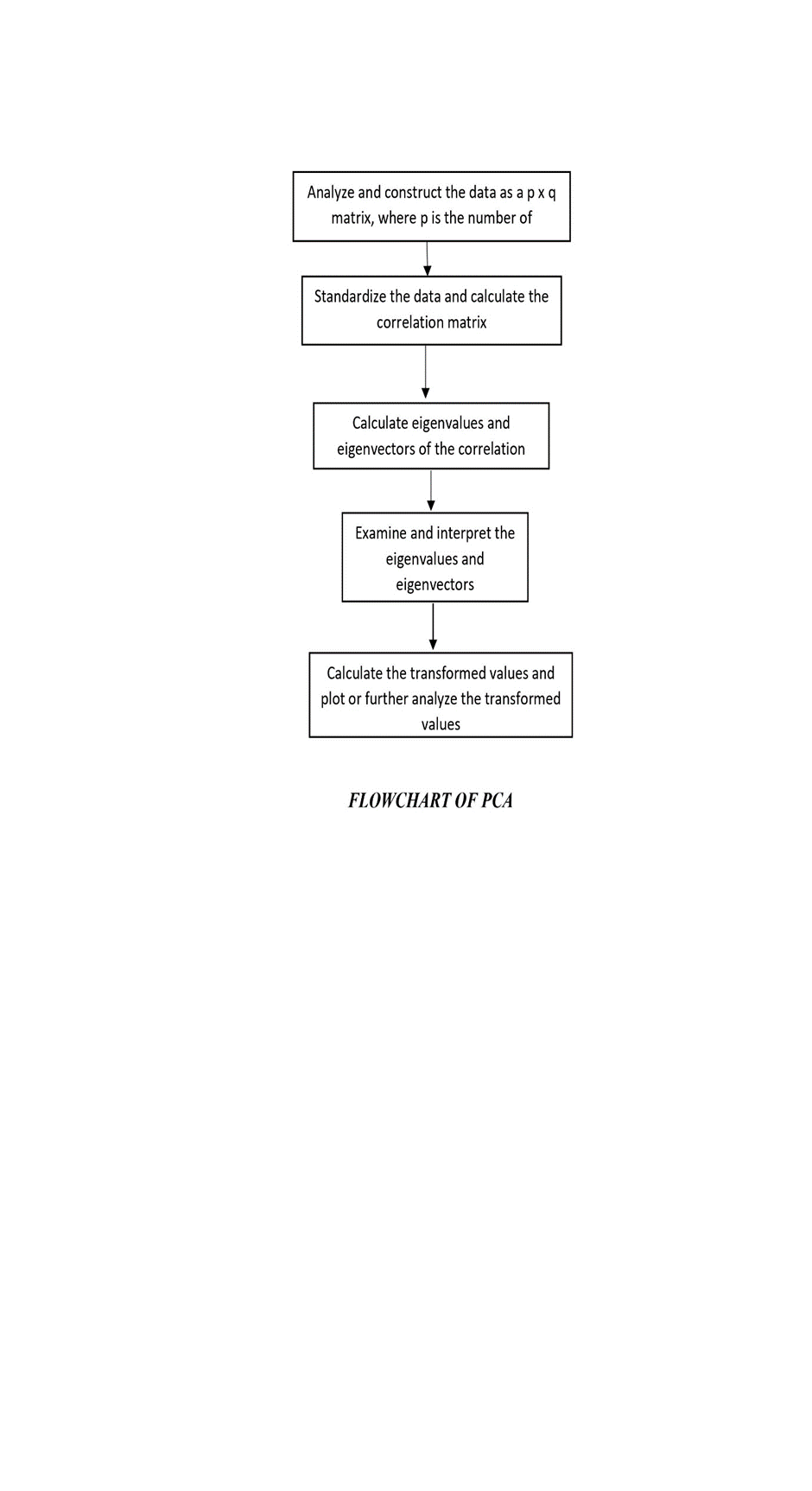
* 1st principal component captures most of the variance in the data.
* 2nd orthogonal to 1st one captures the remaining variance, which is left by 1st component, and so on.

These principal components are uncorrelated and are ordered in such a way that the 1st several components explains most of the variance of the original data. They are uncorrelated because, eigen vectors with distinct eigen values are linearly independent.

We are basically transforming matrices in spaces and

* Eigen vector is the direction that does not change direction.
* Eigen value is the scale of the stretch.

PCA uses the sample covariance matrix to compute rotation, if the rotation matrix is not available at the output, it is proved that an infinite no. of information is lost.



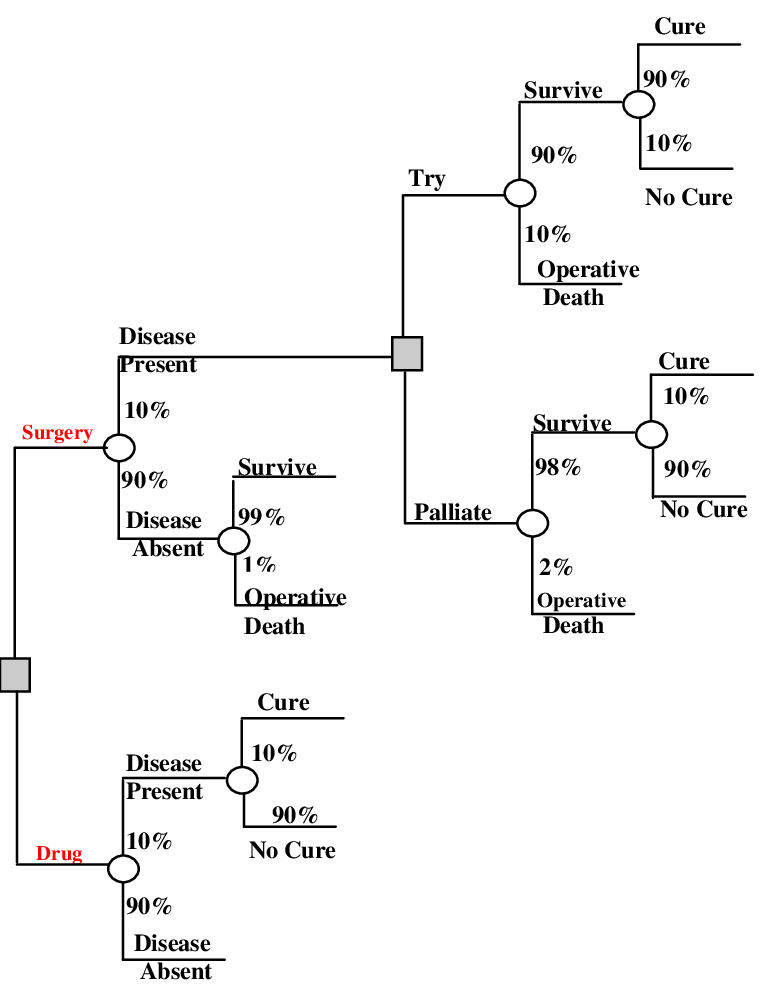
* ***Decision tree (Cancer Type : thyroid, throat, prostate, leukaemia, breast)***

 Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, it’s also widely used in machine learning.

Decision trees create user-friendly rules which makes easy to understand algorithms to the users. Decision trees provide a clear indication of important attributes. Decision trees require less computation compared to other classification algorithms. For example when we are implementing decision trees to detect breast cancer then leaf nodes are divided into two categories: Benign or Malignant. Rules will be established among the chosen data set attributes in order to determine if the tumour is benign or malignant.

We can consider the following example for understanding decision trees.

The following flowchart represents a decision tree deciding if there is a cure possible or not after performing surgery or by prescribing medicines…



* ***SVM (Cancer Type : breast, esophageal, liver, ocular, osteosarcoma, prostate)***

Support Vector Machine (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) which can be used for both classification and regression challenges. However, it is mostly used in classification problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate N-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

So, the basic objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane.

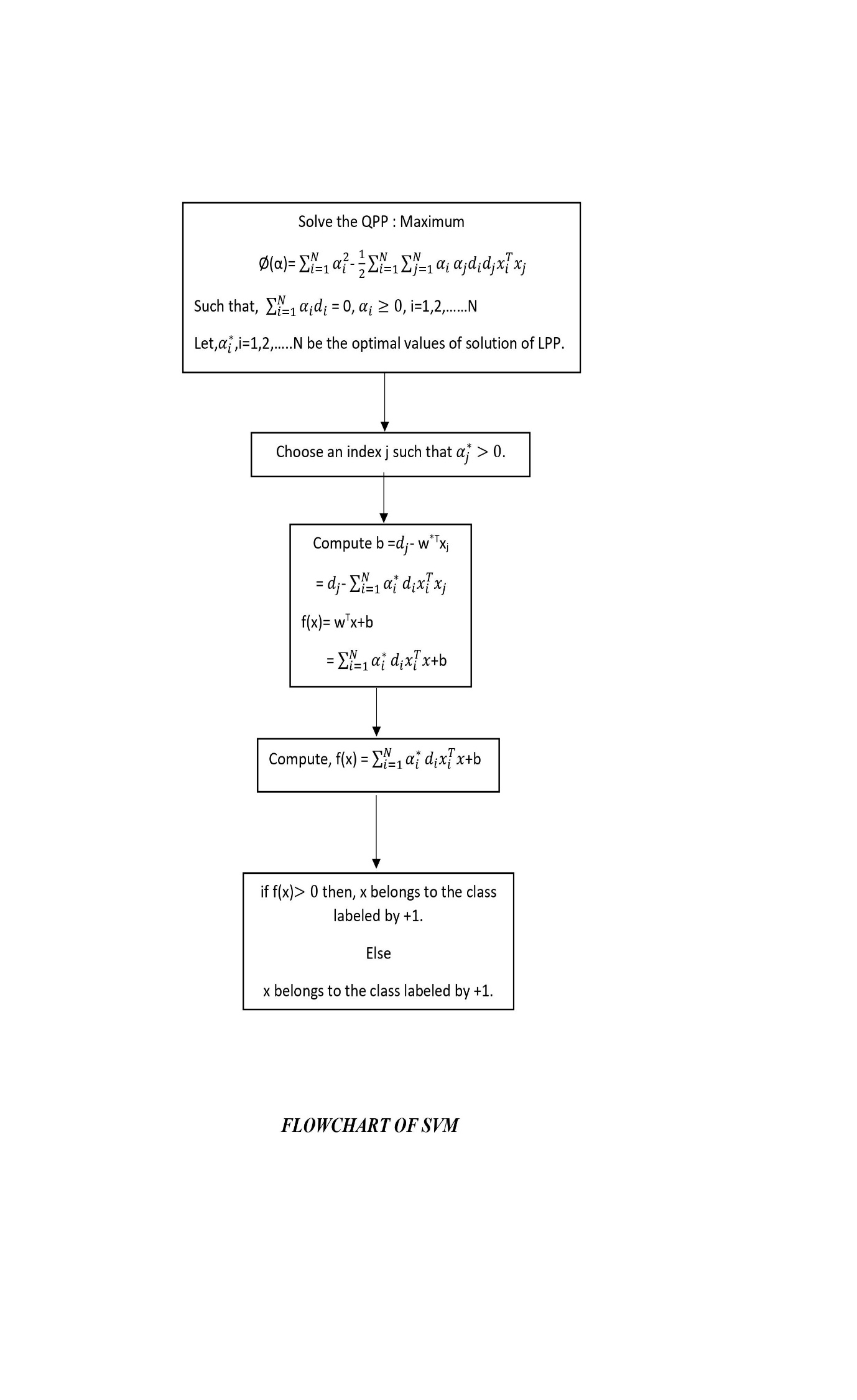
The use of the SVM classifier in the medical diagnosis is increasing gradually, and there is no doubt that evaluation of data taken from patients and decisions of the experts are the most important factors in diagnosis.

SVM can be of two types:

1. **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
2. **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

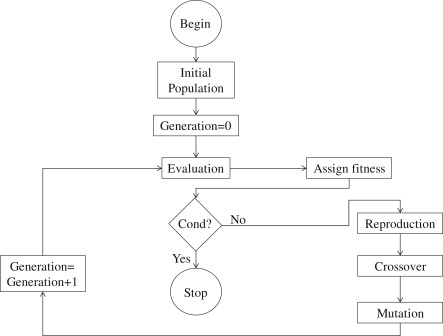
SVM has a strong generalization performance and classification precision compared with other classification approaches. It constructs an effective classifier for cancer diagnosis which can be utilized in real clinical diagnostic system and thereby assist clinical physicians in making correct and effective decisions.

* SVM provides a good effect in colour image segmentation based on pixel classification.
* SVM can reduce the segmentation error which caused by fast movement of the object.
* It decreases the computational error and time and increases the quality of the colour image segmentation.
* SVM is capable for edge detection, region margining, statistical pixel classification.
* SVM partitions an image into regions of connected pixels according to image intensity distribution.



* ***Genetic algorithm (Cancer Type : skin, prostate)***

Genetic algorithms are randomized search algorithms that have been developed in an effort to imitate the mechanics of natural selection and natural genetics. Genetic algorithm is a kind of stochastic algorithm based on the theory of probability. In application this method to a stage wise superstructure model, the search process is determined by stochastic strategy. Genetic algorithms operate on string structures, like biological structures, which are evolving in time according to the rule of survival of the fittest by using a randomized yet structured information exchange. Thus, in every generation, a new set of strings is created, using parts of the fittest members of the old set. The key ideas of these algorithms were the creation of a population of solutions, with some method of varying the composition of the individuals in the population based on a measure of fitness, or how well the individuals perform on a particular task. The most fit individuals created would replace less fit individuals so that over time the overall fitness of the population would increase. Genetic programming has been developed and extensively applied for analysis of molecular data to classify cancer subtypes and characterize the mechanisms of cancer pathogenesis and development.



***FLOWCHART OF GENETIC ALGORITHM***

* ***Naïve Bayes Classifiers (Cancer Type : prostate)***

The Naive Bayes Classifier technique is based

on the so-called Bayesian theorem and is par-

ticularly suited when the dimensionality of the

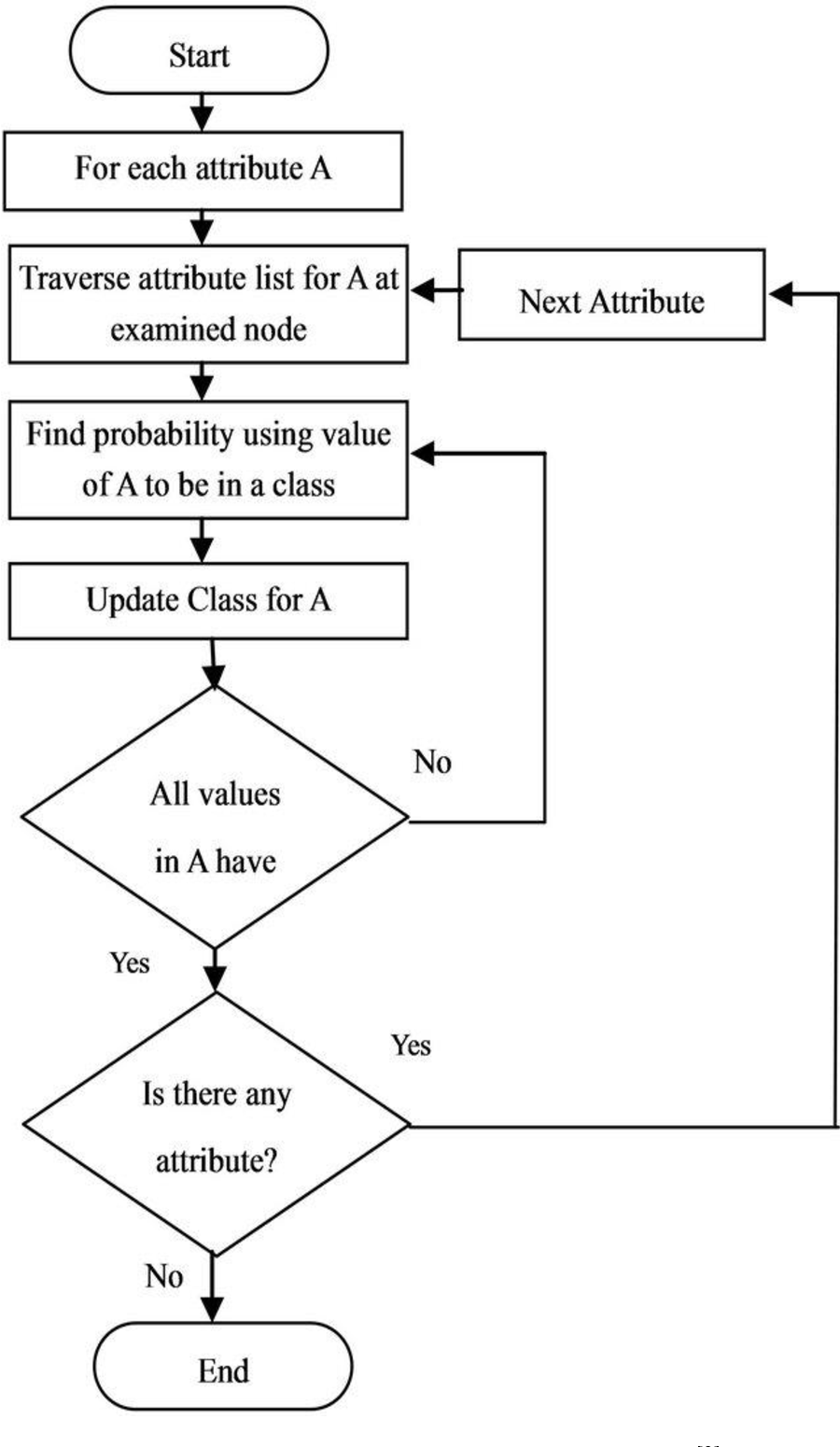
inputs is high. Despite its simplicity, Naive

Bayes can often outperform more sophisticated

classification methods

Classification is supervised learning which maps data into predefined groups or classes .Classification applications includes image and pattern recognition, medical diagnosis ,detecting faults and many more. Accuracy and the interpretability are the two parameters used to find the efficiency of classification model. The Naive Bayes Classifier technique is a supervised learning algorithm, which is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes often used for solving more sophisticated classification methods. Naïve Bayes Classifier is a probabilistic classifier and one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions on the basis of the probability of an object. The prediction for cancer with the help of this technique is performed from mining the patient’s historical data or data repository. Further from the experimental results it has been found that Naive Bayes Classifiers is providing improved accuracy with low computational effort and very high speed.

***FLOWCHART OF THE ABOVE METHOD IS GIVEN BELOW…***



The Naive Bayes Classifier technique is based

on the so-called Bayesian theorem and is par-

ticularly suited when the dimensionality of the

inputs is high. Despite its simplicity, Naive

Bayes can often outperform more sophisticated

classification methods

The Naive Bayes Classifier technique is based

on the so-called Bayesian theorem and is par-

ticularly suited when the dimensionality of the

inputs is high. Despite its simplicity, Naive

Bayes can often outperform more sophisticated

classification methods

The Naive Bayes Classifier technique is based

on the so-called Bayesian theorem and is par-

ticularly suited when the dimensionality of the

inputs is high. Despite its simplicity, Naive

Bayes can often outperform more sophisticated

classification methods

**RECENT DEEP LEARNING TECHNIQUES USED FOR CANCER PREDICTION AND DIAGNOSIS**

* ***CNN***

A Convolutional Neural Network is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. CNN is class of neural networks, most commonly applied to analyse visual imagery. The pre-processing required in a CNN is much lower as compared to other classification algorithms. It is one of the main ML/DL techniques to do images recognition, images classifications. Objects detections, pattern recognition etc., are some of the areas where CNNs are widely used. CNN image classifications takes an input image, process it and classify it under certain categories. Computers sees an input image as array of pixels and it depends on the image resolution.

The CNN architecture has two main types of transformation. The first is convolution, in which pixels are convolved with a filter or kernel. This step provides the dot product between image patch and kernel. The width and height of filters can be set according to the network, and the depth of the filter is the same as the depth of the input. A second important transformation is subsampling and used as per requirement. The size of the pooling filter can be set by the user and is generally taken in odd numbers. The pooling layer is responsible to lower the dimensionality of the data, and is quite useful to reduce over fitting. After using a combination of convolution and pooling layers, the output can be fed to a fully connected layer for efficient classification.

Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1.  The role of the CNN is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction. This is important when we are to design an architecture which is not only good at learning features but also is scalable to massive datasets.

Mathematical Convolution is an operation on two functions that produces a third function that expresses how the shape of one is modified by the other.

We take a KERNEL and apply it on the input image to get the convoluted feature. This feature is passed to the next layer.

CNN are composed of multiple layers (collection of nodes, like for input layer which contains the raw data, each variable is considered as node) of artificial neurons, which are mathematical functions that calculates the weighted sum of multiple inputs and output an activation value.

When we put an input image, each layer generates several activation functions that are passed onto the next layer.

* 1st layer classify the basic feature like horizontal and diagonal edges.
* 2nd layer classify more complex feature like corners or combinatorial edges.

As we move deeper to the network it can even classify the objects or faces.

Based on activation map of the final convolutional layer, the classification layer identifies how likely the image belongs to a “CLASS”.

Pooling layer reduces the spatial size of the convoluted feature, to decrease the computational complexity to process the data. This also works to reduce the noise of the data.

A Hidden layer (called hidden because they are not directly observable from the system inputs and outputs) is a layer between input and output layers, where artificial neurons take in set of weighted input and produce an output through an activation function. This allows a neural network to be broken down into several transformation of the data.

Modern medical image processing techniques work on histopathology images captured by a microscope, and then analyze them by using different algorithms and methods. Machine learning algorithms are now being used for processing medical imagery and pathological tools. Manual detection of a cancer cell is a tiresome task and involves human error, and hence computer-aided mechanisms are applied to obtain better results as compared with manual pathological detection systems. In deep learning, this is generally done by extracting features through a convolutional neural network (CNN) and then classifying using a fully connected network. Deep learning is extensively utilized in the medical imaging field.

# Section 5: Drawbacks/area of improvement in ML

1. ***The basic and general drawbacks of Machine Learning Methods are more or less like the following:-***

* *Possibility of High Error* : Suppose we train an algorithm with data sets small enough to not be inclusive. We will end up with biased predictions coming from a biased training set. Such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it. The main problem occurs in the training and testing of data. The data is huge, so sometimes removing errors becomes nearly impossible.
* *Algorithm Selection* : The selection of an algorithm in Machine Learning is still a manual job. We have to run and test our data in all the algorithms. After that only we can decide what algorithm we want. We choose them on the basis of result accuracy. The process is very much time-consuming. So, one must carefully choose the algorithms for the relevant purpose.
* *Time & Space* : ML needs enough time to let the algorithms learn and develop enough to fulfil their purpose with a considerable amount of accuracy and relevancy. If the data is large and advanced, the system will take time. This may sometimes cause the consumption of more CPU power. Even with GPUs alongside, it sometimes becomes hectic. Also, the data might use more than the allotted space.
* *Data Acquisition* : Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. We take a huge amount of data for training and testing. This process can sometimes cause data inconsistency. The reason is some data constantly keep on updating. There can also be times where they must wait for new data to be generated. If not, the old and new data might give different results which is not a good sign for an algorithm.
* *High Cost Implementations* : Setting up AI-based machines, computers, etc. entails huge costs given the complexity of engineering that goes into building one.
* *Inability to Replace Humans* : It is beyond any doubt that machines perform much more efficiently as compared to a human being. But even then it is practically impossible to replace humans with AIs, at least in the near future, because we can’t build human intelligence in a machine as it is a gift of nature. So, no matter how smart a machine can become, it can never replace a human. One of the most amazing characteristics of human cognitive power is its ability to develop with age and experience. However, the same can’t be said about machines as they can’t improve with experience, rather it starts to wear and tear with time.
* *High Chances of Misuse* : Machines are rational but don’t have any emotions or moral values. Machines can’t judge what is right or what is wrong because they are incapable of understanding the concept of ethical or legal. They are programmed for certain situations and as such can’t take decisions in cases where they encounter an unfamiliar or unfair (not programmed for) situation.
* *Lacks Creativity* : The creativity of machines is limited to the creative ability of the person who programs and commands them. Human brains are characterized by immense sensitivity and high emotional quotient. So, machines can become skilled but they can never acquire the abilities of the human brain. The reason is that skills can be learned and mastered, but abilities come naturally.
* *Risk of Unemployment* : Nowadays, most of the organization are implementing automation at some level in order to replace the minimum qualified individuals with machines that can do the same work with higher efficiency. While rapid development being made in the field of AI, we are not sure whether AIs will lead to higher unemployment or not. But AIs are likely to take over the majority of the repetitive tasks, which are largely binary in nature and involve minimum subjectivity. So, it can’t be ruled out that AIs will result in a less human intervention which may cause major disruption in the employment standards.

(AI : Artificial Intelligence)

1. ***The general drawbacks of Machine Learning Methods particularly in Medical field***

The biggest problem in pathological image analysis using well known machine learning models, is that only a small number of training data is available. Also it is time consuming and need an expert for following the evaluation methodologies depending on whether it is a problem of detection or grading. Moreover, only grading is not enough to conclude the condition of patient. To find the exact stage of the cancer patient, further clinical tests like MRI and PET are needed which are relatively expensive. Perhaps in past research papers, special attention for data quality and integrity, separating training and testing datasets clearly and explaining the evaluation methodology have been overlooked largely due to the lack of computational resources and the relatively high cost of digital imaging equipment for pathology. Moreover, most of the machine learning methods discussed in published research papers may help in determining the outcome for cancer patients at a particular hospital, for a particular time, but they may become irrelevant over time.

* **Table (Drawbacks of mentioned ML techniques for Cancer Prognosis)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the ML Method** | **Learning Type** | **General Cancer Type** | **Drawbacks of the Method** |
| **Fuzzy Logic** | supervised | bladder, breast, throat, lung | Since the rules are based on predetermined rules, the decision-making process is based on fuzzy logic. If these rules are  ﬂawed, the results may not be acceptable at all. Choosing a membership function and basic rules is one of the most  challenging parts of creating fuzzy systems. On the other hand, the implementation of fuzzy logic in standard hardware  Since the rules are based on predetermined rules, the decision-making process is based on fuzzy logic. If these rules are  ﬂawed, the results may not be acceptable at all. Choosing a membership function and basic rules is one of the most  challenging parts of creating fuzzy systems. On the other hand, the implementation of fuzzy logic in standard hardware  Since the rules are based on predetermined rules, the decision-making process is based on fuzzy logic. If these rules are  ﬂawed, the results may not be acceptable at all. Choosing a membership function and basic rules is one of the most  challenging parts of creating fuzzy systems. On the other hand, the implementation of fuzzy logic in standard hardware  Since the rules are based on predetermined rules, the decision-making process is based on fuzzy logic. If these rules are  ﬂawed, the results may not be acceptable at all. Choosing a membership function and basic rules is one of the most  challenging parts of creating fuzzy systems. On the other hand, the implementation of fuzzy logic in standard hardware  Since the rules are based on predetermined rules, the decision-making process is based on fuzzy logic. If these rules are  ﬂawed, the results may not be acceptable at all. Choosing a membership function and basic rules is one of the most  challenging parts of creating fuzzy systems. On the other hand, the implementation of fuzzy logic in standard hardware  Since the rules are based on predetermined rules, the decision-making process is based on fuzzy logic. If these rules are flawed, the results may not be acceptable at all. Choosing a membership function and basic rules is one of the most challenging parts of creating fuzzy systems. On the other hand, the implementation of fuzzy logic in standard hardware requires multiple and time-consuming experiments and the efficiency of fuzzy logic in recognizing the pattern is less than the neural network in machine learning which it is less covered in Data Science. Moreover, Use of fuzzy logic may become an obstacle to the verification of system reliability and Requires tuning of membership functions. Fuzzy Logic may not scale well to large or complex problems. It deals with imprecision but not uncertainty. |
| **ANN** | unsupervised | bladder, breast, brain, cervical, colorectal, esophageal, liver, lung, lymphoma, head/neck, prostate, skin, stomach, throat, thoracic | ANNs suffer from the limited capability to explicitly explain the causal relationships between predictive factors and outcomes, which is probably the most major drawback. Another drawback is that a well trained model would be difficult to share with other researchers. This is because the knowledge discovered from the data is all encoded into a huge weight matrix, which is difficult to interpret and share. Furthermore, the complexity of the model structure in ANNs makes it more prone to overfitting, the case where the network overlearns and mimics the training dataset but performs poorly when presented to an external dataset. |
| **K-Means clustering** | supervised | pleural mesothelioma, breast, colorectal, lymphoma | Choosing the number of clusters k can be difficult even if we have a static data set and previous domain knowledge about the data. Also, k-means can only handle numerical data and the results might be skewed if we do not normalize it. |
| **PCA** | supervised | pleural mesothelioma, breast, colorectal, lymphoma | After implementing PCA on the dataset, our original features will turn into Principal Components. Principal Components are the linear combination of our original features. Principal Components are not as readable and interpretable as original features. Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features. Standardizing data before implementing PCA is must, otherwise PCA will not be able to find the optimal Principal Components. |
| **Decision Tree** | supervised | thyroid, throat, prostate, leukaemia, breast | 1. The mathematical calculation of decision tree mostly require more time and memory. 2. The reproducibility of decision tree model is highly sensitive as small change in the data can result in large change in the tree structure. 3. The space and time complexity of decision tree model is relatively higher. 4. Decision tree model training time is relatively more as complexity is high. 5. Single Decision tree is often a weak learner so we require a bunch of decision tree for called random forest for better prediction. |
| **SVM** | supervised | ***breast, esophageal, liver, ocular, osteosarcoma, prostate*** | 1. SVM algorithm is not suitable for large data sets. SVM takes a long training time on large datasets. 2. Algorithmic complexity and memory requirements of SVM are very high. Need a lot of memory since we have to store all the support vectors in the memory and this number grows abruptly with the training dataset size. 3. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping. 4. One must do feature scaling of variables before applying SVM. 5. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform. 6. As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification. 7. Choosing an appropriate Kernel function (to handle the non-linear data) is not an easy task. It could be tricky and complex. In case of using a high dimension Kernel, we might generate too many support vectors which reduce the training speed drastically. 8. SVM model is difficult to understand and interpret by human beings unlike Decision Trees. |
| **Genetic Algorithm** | supervised | ***skin, prostate*** | This method requires less information about the problem, but designing an objective function and getting the representation and operators right can be difficult. Also this is computationally expensive i.e. time-consuming. |
| **Naïve Bayes Classifiers** | supervised | ***prostate*** | 1. Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.  2. If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation. |
| **CNN** | **Deep learning** | **breast, lung, skin, prostate** | Since convolutional neural networks are typically used for image-classification, we are generally dealing with high-dimensional data (images). While the structure of a ConvNet aims to mitigate over-fitting, it generally requires a large dataset to process and train the neural network effectively.  If the CNN has several layers then the training process takes a lot of time if the computer doesn’t consist of a good GPU. |

# Section 6: Conclusion

From this paper we can fully infer that machine learning plays a very important role in a sensitive medical subject like cancer detection. At the same time we also see that machine learning has many small and big difficulties in risky things like cancer prognosis. But the relief is that we are constantly trying to overcome these difficulties and we have been able to improve a lot. Despite the disadvantages of the various machine learning methods discussed here, there is no hesitation in acknowledging the fact that these methods have been widely accepted and used in various fields of cancer detection in the past due to their problem solving skills in less time, accuracy of result obtained, ability to reduce human labour a lot, ability to predict correct result.

Here we have tried to give a fair idea about the working methods of so-called machine learning techniques, the acceptance of which techniques are relatively high in the case of any cancer, and their shortcomings.

By analyzing the advantages and disadvantages of these methods in the medical field together, it can be said that if the working methods of these techniques can be made more advanced or improved more practically and more efficiently in the future, then there is no doubt in accepting that, in future machine learning will surely take the main role in the most relevant subject like Cancer Detection and Prognosis. The biggest application of this automated cancer diagnosis will be therapy prediction, personalized medicine and prognosis.

This will solve the problems that go beyond just diagnosis and it will give a new direction to human civilization and strength to lead the way of life.

# References

1. AlexanderSelvikvågLundervolda, b. A. (n.d.). *An overview of deep learning in medical imaging focusing on MRI.*

2. Berkman Sahiner, a. a. (n.d.). Deep learning in medical imaging and radiation therapy.

3. Bradley J. Erickson, P. K. (n.d.). Machine Learning for Medical Imaging1.

4. Chetan L. Srinidhia, O. C. (n.d.). *Deep neural network models for computational histopathology: A survey.*

5. Daisuke Komura, S. I. (n.d.). *Machine learning methods for histopathological image analysis.*

6. *Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours.* (n.d.). www.nature.com/scientificreports.

7. Geng-Shen Fu, 1. Y.-S.-H. (n.d.). Machine Learning for Medical Imaging.

8. Jahanzaib Latif, A. I. (n.d.). Medical Imaging using Machine Learning and Deep Learning Algorithms: A Review.

9. Joseph A. Cruz, D. S. (n.d.). *Applications of Machine Learning in Cancer Prediction.*

10. Kan, A. (n.d.). Machine Learning Applications in Cell Image Analysis. *Immunology and Cell Biology*.

11. Kaustav Bera, M., Ian Katz, M., & and Anant Madabhushi, P. (n.d.). Reimagining T Staging Through Artificial Intelligence and Machine Learning Image Processing Approaches in Digital Pathology.

12. Khushboo Munir 1, \*. H. (n.d.). Cancer Diagnosis Using Deep Learning: A Bibliographic Review.

13. Liao, S. C. (n.d.). Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations.

14. Majid Masoumi, S. H. (n.d.). THE CHALLENGES AND ADVANTAGES OF FUZZY SYSTEMS APPLICATIONS.

15. Megha Rathi\*, A. K. (n.d.). Breast Cancer Prediction using Naïve Bayes Classifier.

16. Muhammad Imran Razzak, S. N. (n.d.). *Deep Learning for Medical Image Processing: Overview, Challenges and Future.*

17. Naira Elazab 1, H. S.-S. (n.d.). Objective Diagnosis for Histopathological Images Based on Machine Learning Techniques: Classical Approaches and New Trends.

18. Tabassum Yesmin Rahman1, L. B. (n.d.). *Automated Oral Squamous Cell Carcinoma Identification using Shape, Texture and Color features of Whole Image Strips*.

19. Tairan Liu1, 2. K. (n.d.). Deep learning-based holographic polarization microscopy.

20. Turgay Ayer, 1. Q. (n.d.). Artificial Neural Networks in Mammography Interpretation and Diagnostic Decision Making.

21. Vadim Zinchuk1, 3. a.-Z. (n.d.). Machine Learning for Analysis of Microscopy Images: A Practical Guide.