**CHUKA** **UNIVERSITY.**

**NAME: JANET NASHIPAI.**

**REG: EB3/43467/19.**

**UNIT: ACSC 482 SEMINARS IN APPLIED COMPUTER SCIENCE.**

**COURSE: APPLIED COMPUTER SCIENCE.**

**NAME OF SUPERVISING LECTURER: MR. KEVIN GOGO.**

**RESEARCH TOPIC: CANCER PREDICTION USING MACHINE LEARNING.**

**RESEARCH SUBMITED TO CHUKA UNIVERSITY COMPUTER SCIENCE DEPARTMENT, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE IN APPLIED COMPUTER SCIENCE.**

**Submitted on………………………………………**

# **DECLARATION**

I hereby declare that this Proposal is my own work and has, to the best of my knowledge, not been submitted to any other institution of higher learning.

Student: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Registration Number: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: ............................................... Date: .....................................................

This proposal has been submitted as a partial fulfillment of requirements for the Bachelor of Science in Applied Computer Science of Chuka University with my approval as the University supervisor.

Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: ..................................................... Date: ..................................................

# **ACKNOWLEDGEMENT.**

Personally, I would like to acknowledge the following persons for their help and support in this project:

* My supervisor, Mr. Kevin Gogo from the department of Computer Science, Chuka University for his proper guidance and direction till its completion.
* My parents for paying my school fees and ensuring that I was in school at all time until the project came to an end.
* And above all, I would like to thank the Almighty for giving me life and enabling me to do this research.

**ABSTRACT**.

This document is a five chapter document which describes the introduction of the research study including the statement of the problem and defining the objectives together with justification and the scope of the research. It also has the literature search of the study which was done extensively by reviewing existing literature and articles done by other researchers about the research problem. In addition is the methodology which describes methods of collecting data and how the research objectives can be achieved. This document also has results, discussions and findings, future work and challenges together with conclusions about the research.

Machine learning, a branch of artificial intelligence employs statistical, probabilistic, optimization technique that allows the computer to learn from examples and to detect patterns from the data sets. This ability is very well suited to medical application especially the ones that depend on genomic measurements. In this case, machine learning is very often used cancer diagnosis and detection. Recently, it has been applied in cancer prediction. In the study, I conducted survey of the types of machine learning methods being used, type of data integrated and performance of these methods of cancer prediction. A few trends are noted, which include protein biomarker growing dependency and microarray data, reliance on older technologies like artificial neural networks (ANNs) instead of recently developed and easily interpretable machine learning methods. A number of conducted studies also appear to lack appropriate testing levels. Among the better studies, its clear that machine learning methods can be used(15-25%) to improve the accuracy of predicting cancer susceptibility, recurrence and mortality. At a higher level, it is evident that machine learning greatly helps in our understanding of cancer development and progression.

**ACRONYMS**

ML- Machine Learning

ANN- Artificial Neural Network

SVM- Support Vector Machine

DT- Decision Tree

BN- Bayesian Network

SSL- Semi-supervised Learning

TCGA- The Cancer Genome Atlas Research Network

HTT- High-throughput Technologies

OSCC- Oral Squamous Cell Carcinoma

CFS- Correlation based Feature Selection

AUC-Area Under Curve

ROC-Receiver Operating Characteristic

BCRSVM-Breast Cancer Support Vector Machine

PPI- Protein–Protein Interaction

GEO- Gene Expression Omnibus

LCS- Learning Classifying Systems

ES-Early Stopping algorithm

SEER- Surveillance, Epidemiology and End results Database

NSCLC- Non-small Cell Lung Cancer

NCI caArray- National Cancer Institute Array Data Management System

**Table of Contents**

[**DECLARATION** ii](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458402)

[**ACKNOWLEDEGMET.** ii](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458403)

[**ABSTRACT.** iii](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458404)

[**List of Tables and Figures.** 6](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458405)

[**ACRONYMS.** 6](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458406)

[**CHAPTER 1: INTRODUCTION.** 8](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458407)

**1.1: Background of the study**………………………………………………………………………………………………………

[**1.2: Problem Statement.** 9](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458409)

[**1.3: Objectives.** 9](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458410)

[**1.4: Scope.** 9](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458411)

[**1.5: Justification.** 9](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458412)

[**CHAPTER 2: LITERATURE REVIEW.** 11](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458413)

[**2.1: Introduction.** 11](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458414)

[**2.2: Literature Search.** 11](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458415)

[**CHAPTER 3: RESEARCH METHODOLOGY.** 21](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458416)

[**3.0: Introduction.** 21](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458417)

[**3.1: Research Design.** 21](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458418)

[**3.2: Research Tools and Procedures.** 22](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458419)

[**3.2.1: Data Collection and analysis.** 23](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458420)

[**3.2.2: Study.** 25](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458421)

[**3.2.3: Interview and Observation.** 27](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458422)

[**3.2.4: Datasets.** 28](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458423)

[**3.2.5: Classifier.** 29](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458424)

[**CHAPTER 4: RESULTS AND DISCUSSION.** 38](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458430)

[**CHAPTER 5: CONCLUSION, RECOMMENDATIONS, CHALLENGES AND FUTURE WORK.** 40](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458431)

[**5.1: Conclusion.** 40](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458432)

[**5.2: Recommendation.** 40](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458433)

[**5.3: Challenges and Future Work.** 41](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458434)

[**REFERENCES.** 43](file:///C:\Users\JANET\AppData\Local\Microsoft\Windows\INetCache\IE\2EL4WZG2\Kirwa's_Seminar_Final%5b1%5d.docx#_Toc118458435)

**CHAPTER ONE: INTRODUCTION.**

**1.1:Background of the study.**

Cancer is one of the most death threatening disease across the world today. It is a condition where cells lose its ability to control its growth. There are different types of cancer depending on the tissue affected. Cells that cause cancer occur as a series of mutation. Research has proven that the root cause are mutations or abnormal change in the genetic material- deoxyribonucleic acid (DNA).It is also noted that a single mutation cannot cause cancer, however, the accumulation of the mutation on the major part of genome will cause cancer.

In the 3rd world countries, cancer is the 3rd leading cause of death and it is responsible for 7% annual deaths. Basing on a statistical data, it has been estimated that from 2012-2018, the annual incidence of cancer increased from 37,000 to 47,887. During the same period, annual cancer mortality rate rose to 16%.Most common cancer affecting Kenyans are breast, cervical, prostate, esophageal and colorectal and the leading being breast cancer.

However, doctors use some of these approaches to diagnose cancer that is, physical exams where they feel areas of the body for lumps or change in skin color or even enlargement of an organ. Another method is laboratory test where they test the urine and blood. This helps the doctor identify abnormalities that can cause cancer. Imaging tests is another way of diagnosing cancer where the doctor examines the bones and internal organs by using computerized tomography (CT)scan, magnetic resonance imaging (MRI),position emission tomography(PET)scan, ultrasound and x-ray, among others. Lastly the doctors may use biopsy where they collect the samples of the cells for testing. The biopsy procedure depends on the type of cancer and its location. These methods have been used by the doctors for many years,

Although these techniques have been effective, in some cases some cancer can be hard to spot than others. Some forms of disease are rare hence doctors paying less attention to them. Less is invested in learning on how to diagnose and treat them. Sometimes, even with more common cancer, figuring out what’s cancer and what’s isn’t can be difficult. It can be difficult to find cancerous cells, or to figure out if a cell or a group of cells is cancerous.

Therefore, to prevent the above stated disadvantages that can lead to lose of many lives, machine learning algorithms can be used in prediction and prognosis of cancer. Some of these algorithms may include K-nested neighbor, support vector machine, artificial neural network, among others. Almost all these machine learning algorithm employ supervised learning. These predictions are made using three types of input data: genome data (single-nucleotide polymorphism, mutations, microarrays), proteomic data (specific protein biomarkers, mass spectral analysis) clinical data (histology, tumor staging, tumor size, age, weight, risk behavior).

**1.2:Problem Statement**.

Cancer is one of the main causes of death world-wide. Early diagnostics have increased the chances of correct treatments and survival. However, we find this process tedious and frequently leads to disagreement between pathologists.

Computer-aided diagnosis showed the potential of improving diagnostic accuracy. But early detection and prevention can greatly reduce the chances of death. It is important to detect cancer as early as possible. Some of the gaps identified during the study were: how to implement sustainable lifestyle changes, understanding molecular mechanism of tumor, developing interventions or support to improve survivorship experience of the patient.

To overcome these problems, we are using some of the machine learning algorithm which will help us to predict cancerous cells based on data present in the algorithm. By using these algorithms, we can predict cancer at early stages and proper medication process can be conducted and lives will be saved.

**1.3: Objectives**.

The main objective of this research is to predict cancer using machine learning. Other objectives include:

1. To produce a model that can be used to perform classification, prediction, estimation or any other similar task by the help of machine learning technique
2. To train datasets using machine learning algorithms.
3. To improve the accuracy of cancer susceptibility, recurrence and mortality
4. To discover, identify patterns and relationships between the data sets using machine learning algorithm

**1.4: Scope.**

The study of this research can be adopted in predicting cancer using machine learning algorithms in order to save lives and for proper medication to be conducted. We are using machine learning algorithms to train the datasets and by providing models that can perform classification, prediction and estimation.

**1.5: Justification**

The study will help you to predict cancer in their early stages. In addition, this study helps to improve the accuracy of cancer susceptibility, recurrence and mortality.The significance of this study is that machine learning technique can help to produce a model that can be used to perform classification, prediction, estimation or any other similar task.

**CHAPTER TWO: LITERATURE REVIEW.**

**2.1: Introduction.**

In this chapter, a review of detailed information about cancer, machine learning and how to predict cancer in general, how it was done before, currently and the future works has been presented.

**2.2: Literature Search**.

A continual evolution in cancer research has been carried out during the past several years. In order to identify certain cancer forms before they produce symptoms, researchers have used a variety of techniques, such as early-stage screening. Additionally, they have developed novel techniques for early cancer treatment outcome prediction. Large volumes of cancer data have been gathered and are available to the medical research community thanks to the development of new technologies in the field of medicine. However, one of the most fascinating and difficult problems for doctors is making an accurate prediction of a disease's course. As a result, ML techniques are now widely used by scientists conducting research in medicine. While these are complicated datasets, these techniques can find and uncover patterns and correlations between them.

Given the significance of personalized medicine and the growing trend on the application of ML techniques, we here present a review of studies that make use of these methods regarding the cancer prediction and prognosis. In the study, predictive features are considered which may be independent of a certain treatment or are integrated in order to guide therapy for cancer patients, respectively. Moreover, we discuss the types of ML methods being used, the types of data they integrate, the overall performance of each proposed scheme while we also discuss their pros and cons.

An obvious trend in the proposed works includes the integration of mixed data, such as clinical and genomic. However, a common problem that we noticed in several works is the lack of external validation or testing regarding the predictive performance of their models. It is clear that the application of ML methods could improve the accuracy of cancer susceptibility, recurrence and survival prediction. Based on the accuracy of cancer prediction outcome has significantly improved by 15%–20% the last years, with the application of ML techniques.

Several studies have been reported in the literature and are based on different strategies that could enable the early cancer diagnosis. Specifically, these studies describe approaches related to the profiling of circulating miRNAs that have been proven a promising class for cancer detection and identification. However, these methods suffer from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumors. These studies list the potential as well as the limitations of [microarrays](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/microarrays) for the prediction of cancer outcome. Even though gene signatures could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed.

In the present work only studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented. ML, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference. Every learning process consists of two phases: (i) estimation of unknown dependencies in a system from a given dataset and (ii) use of estimated dependencies to predict new outputs of the system. ML has also been proven an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an n-dimensional space for a given set of biological samples, using different techniques and algorithms. There are two main common types of ML methods known as supervised learning and unsupervised learning. In supervised learning a labeled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labeled examples are provided and there is no notion of the output during the learning process. As a result, it is up to the learning model to find patterns or discover the groups of the input data. In supervised learning this procedure can be thought as a classification problem. The task of classification refers to a learning process that categorizes the data into a set of finite classes. Two other common ML tasks are regression and clustering. In the case of regression problems, a learning function maps the data into a real-value variable. Subsequently, for each new sample the value of a predictive variable can be estimated, based on this process. Clustering is a common unsupervised task in which one tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share.

When applying a ML method, data samples constitute the basic components. Every sample is described with several features and every feature consists of different types of values. Furthermore, knowing in advance the specific type of data being used allows the right selection of tools and techniques that can be used for their analysis. Some data-related issues refer to the quality of the data and the preprocessing steps to make them more suitable for ML. Data quality issues include the presence of noise, outliers, missing or duplicate data and data that is biased-unrepresentative. When improving the data quality, typically the quality of the resulting analysis is also improved. In addition, in order to make the raw data more suitable for further analysis, preprocessing steps should be applied that focus on the modification of the data. A number of different techniques and strategies exist, relevant to data preprocessing that focus on modifying the data for better fitting in a specific ML method. Among these techniques some of the most important approaches include (i) dimensionality reduction (ii) feature selection and (iii) feature extraction. There are many benefits regarding the dimensionality reduction when the datasets have a large number of features. ML algorithms work better when the dimensionality is lower. Additionally, the reduction of dimensionality can eliminate irrelevant features, reduce noise and can produce more robust learning models due to the involvement of fewer features. In general, the dimensionality reduction by selecting new features which are a subset of the old ones is known as feature selection. Three main approaches exist for feature selection namely embedded, filter and wrapper approaches. In the case of feature extraction, a new set of features can be created from the initial set that captures all the significant information in a dataset. The creation of new sets of features allows for gathering the described benefits of dimensionality reduction.

However, the application of feature selection techniques may result in specific fluctuations concerning the creation of predictive feature lists. Several studies in the literature discuss the phenomenon of lack of agreement between the predictive gene lists discovered by different groups, the need of thousands of samples in order to achieve the desired outcomes, the lack of biological interpretation of predictive signatures and the dangers of information leak recorded in published studies.

The main objective of ML techniques is to produce a model which can be used to perform classification, prediction, estimation or any other similar task. The most common task in learning process is classification. As mentioned previously, this learning function classifies the data item into one of several predefined classes. When a classification model is developed, by means of ML techniques, training and generalization errors can be produced. The former refers to misclassification errors on the training data while the latter on the expected errors on testing data. A good classification model should fit the training set well and accurately classify all the instances. If the test error rates of a model begin to increase even though the training error rates decrease then the phenomenon of model overfitting occurs. This situation is related to model complexity meaning that the training errors of a model can be reduced if the model complexity increases. Obviously, the ideal complexity of a model not susceptible to overfitting is the one that produces the lowest generalization error. A formal method for analyzing the expected generalization error of a learning algorithm is the bias–variance decomposition. The bias component of a particular learning algorithm measures the error rate of that algorithm. Additionally, a second source of error over all possible training sets of given size and all possible test sets is called variance of the learning method. The overall expected error of a classification model is constituted of the sum of bias and variance, namely the bias–variance decomposition.

Once a classification model is obtained using one or more ML techniques, it is important to estimate classifier’s performance. The performance analysis of each proposed model is measured in terms of sensitivity, specificity, accuracy and area under the curve (AUC). Sensitivity is defined as the proportion of true positives that are correctly observed by the classifier, whereas specificity is given by the proportion of true negatives that are correctly identified. The quantitative metrics of accuracy and AUC are used for assessing the overall performance of a classifier. Specifically, accuracy is a measure related to the total number of correct predictions.

The predictive accuracy of a model is computed from testing set which provides an estimation of the generalization errors. In order to obtain reliable results regarding the predicting performance of a model, training and testing samples should be sufficiently large and independent

The predictive accuracy of the model is computed from the testing set which provides an while the labels of the testing sets should be known. Among the most commonly used methods for evaluating the performance of a classifier by splitting the initial labeled data into subsets are: (i) Holdout Method, (ii) Random Sampling, (iii) Cross-Validation and (iv) Bootstrap. In the Holdout method, the data samples are partitioned into two separate sets, namely the training and the test sets. A classification model is then generated from the training set while its performance is estimated on the test set. Random sampling is a similar approach to the Holdout method. In this case, in order to better estimate the accuracy, the Holdout method is repeated several times, choosing the training and test instances randomly. In the third approach, namely cross-validation, each sample is used the same number of times for training and only once for testing. As a result, the original data set is covered successfully both in the training and in the test set. The accuracy results are calculated as the average of all different validation cycles. In the last approach, bootstrap, the samples are separated with replacement into training and test sets, i.e. they are placed again into the entire data set after they have been chosen for training.

When the data are preprocessed and we have defined the kind of learning task, a list of ML methods including (i) ANNs, (ii) DTs, (iii) SVMs and (iv) BNs is available. Based on the intension of this review paper, we will refer only to these ML techniques that have been applied widely in the literature for the case study of cancer prediction and prognosis. We identify the trends regarding the types of ML methods that are used, the types of data that are integrated as well as the evaluation methods employed for assessing the overall performance of the methods used for cancer prediction or disease outcomes.

ANNs handle a variety of classification or pattern recognition problems. They are trained to generate an output as a combination between the input variables. Multiple hidden layers that represent the neural connections mathematically are typically used for this process. Even though ANNs serve as a gold standard method in several classification tasks they suffer from certain drawbacks. Their generic layered structure proves to be time-consuming while it can lead to very poor performance. Additionally, this specific technique is characterized as a “black-box” technology. Trying to find out how it performs the classification process or why an ANN did not work is almost impossible to detect. 

DTs follow a tree-structured classification scheme where the nodes represent the input variables and the leaves correspond to decision outcomes. DTs are one of the earliest and most prominent ML methods that have been widely applied for classification purposes. Based on the architecture of the DTs, they are simple to interpret and “quick” to learn. When traversing the tree for the classification of a new sample we are able to conjecture about its class. The decisions resulted from their specific architecture allow for adequate reasoning which makes them an appealing technique.

SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained for SVM. SVM can be used to classify tumors among benign and malignant based on their size and patients' age. The identified hyperplane can be thought as a decision boundary between the two clusters. Obviously, the existence of a decision boundary allows for the detection of any misclassification produced by the method.

BN classifiers produce probability estimations rather than predictions. As their name reveals, they are used to represent knowledge coupled with probabilistic dependencies among the variables of interest via a directed acyclic graph. BNs have been applied widely to several classification tasks as well as for knowledge representation and reasoning purposes.

The last two decades a variety of different ML techniques and feature selection algorithms have been widely applied to disease prediction. Most of these works employ ML methods for modeling the progression of cancer and identify informative factors that are utilized afterwards in a classification scheme. Furthermore, in almost all the studies gene expression profiles, clinical variables as well as histological parameters are encompassed in a complementary manner in order to be fed as input to the prognostic procedure. A research that depicts the distribution in published papers using ML techniques to predict (i) cancer susceptibility, (ii) recurrence and (iii) survival.

The success of a disease prognosis is undoubtedly dependent on the quality of a medical diagnosis; however, a prognostic prediction should take into account more than a simple diagnostic decision. When dealing with cancer prognosis/prediction one is concerned with three predictive tasks: (i) the prediction of cancer susceptibility (risk assessment), (ii) the prediction of cancer recurrence/local control and (iii) the prediction of cancer survival. In the first two cases one is trying to find (i) the likelihood of developing a type of cancer and (ii) the likelihood of redeveloping a type of cancer after complete or partial remission. In the last case, the prediction of a survival outcome such as disease-specific or overall survival after cancer diagnosis or treatment is the main objective. The prediction of cancer outcome usually refers to the cases of (i) life expectancy, (ii) survivability, (iii) progression and (iv) treatment sensitivity.

Major types of ML techniques including ANNs and DTs have been used for nearly three decades in cancer detection . According to the recent PubMed results regarding the subject of ML and cancer more than 7510 articles have been published until today. The vast majority of these publications makes use of one or more ML algorithms and integrates data from heterogeneous sources for the detection of tumors as well as for the prediction/prognosis of a cancer type. A growing trend is noted the last decade in the use of other supervised learning techniques, namely SVMs and BNs, towards cancer prediction. All of these [classification algorithms](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/classification-algorithm) have been widely used in a wide range of problems posed in cancer research.

In the past, the typical information used by the physicians conclude with a reasonable decision regarding cancer prognosis and included histological, clinical and population-based data.  The integration of features such as family history, age, diet, weight, high-risk habits and exposure to environmental carcinogens play a critical role in predicting the development of cancer. Even though this type of macro-scale information referred to a small number of variables so that standard statistical methods could be used for prediction purposes, however these types of parameters do not provide sufficient information for making robust decisions. With the rapid advent of genomic, [proteomic](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/proteomics) and imaging technologies a new kind of molecular information can be obtained. Molecular biomarkers, [cellular parameters](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/cellular-parameters) as well as the expression of certain genes have been proven as very informative indicators for cancer prediction. The presence of such High Throughput Technologies (HTTs) nowadays has produced huge amounts of cancer data that are collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type. Additionally, feature selection methods have been published in the literature with their application in cancer. The proposed computational tools aim at identifying informative features for accurately identification of disease class.

There are nowadays separate subgroups among the same type of cancer based on specific genetic defects that have different treatment approaches and options as well as different clinical outcomes. This is the foundation of the individualized treatment approach, in which computational techniques could help by identifying less costly and effectively such small groups of patients. Furthermore, the development of a community resource project, namely The Cancer Genome Atlas Research Network (TCGA) has the potential support for personal medicine as it provides large scale genomic data about specific tumor types. TCGA provides with the ability to better understand the molecular basis of cancer through the application of high-throughput genome technologies.

An extensive search was conducted relevant to the use of ML techniques in cancer susceptibility, recurrence and survivability prediction. Two electronic databases were accessed namely PubMed, [Scopus](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/scopus). Due to the vast number of articles returned by the search queries, further scrutinization was needed in order to maintain the most relevant articles. The relevance of each publication was assessed based on the keywords of the three predictive tasks found in their titles and abstracts. The majority of these studies use different types of input data: genomic, clinical, histological, imaging, demographic, [epidemiological data](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/epidemiological-data) or combination of these. Papers that focus on the prediction of cancer development by means of conventional statistical methods (e.g. chi-square, Cox regression) were excluded as were papers that use techniques for [tumor classification](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/tumor-classification) or identification of predictive factors. According to their survey based on ML applications in cancer prediction, we noted a rapid increase in papers that have been published in the last decade. Although it is impossible to achieve a complete coverage of the literature, we believe that a significant number of relevant papers were extracted and are presented in this review. As mentioned above, from the initial group of papers we selected a representative list that follows a well-organized structure. Specifically, we selected these studies that make use of recognizable ML techniques and integrated data from heterogeneous sources in order to predict the desirable outcome. We focused mainly on studies that have been published the last 5 years as an aim to present the most recent state of the art in the field and their advances in comparison to older publications. In their research the table data consisted of; Cancer type, ML method, number of patients, type of data as well as the overall accuracy achieved by each proposed method are presented. Moreover, sub-table which corresponds to studies regarding a specific scenario (i.e. cancer susceptibility prediction, cancer recurrence prediction and cancer survival prediction). It should be noted that in articles that more than one ML techniques are applied for prediction.

A detailed analysis of more recent studies revealed that there is a growing trend in risk assessment as well as the prediction of recurrence of a cancer type regardless the ML technique used. Many research groups have tried to predict the possibility of redeveloping cancer after remission and appeared to improve the accuracy of predictions compared to alternative statistical techniques. Moreover, the vast majority of these publications used molecular and clinical data in order to make their predictions. The use of such measurable features as input data is a growing trend based on the advent of HTTs.

In the following, we are going to discuss one case for each of the objectives of predicting (i) susceptibility, (ii) recurrence and (iii) survival, all by means of ML techniques. Each sub-section summarizes the representative studies we have selected based on their predictive outcomes. I only selected those publications that have been accepted the last 5 years and make use of distinguishable ML.

Among the most recent publications that resulted after search regarding the cancer risk assessment prediction, I selected a recent and very interesting study to present relevant to the breast cancer risk estimation by means of ANNs . It is a different study among the others presented in this review article regarding the data type used. Although all of the publications selected make use of molecular, clinical or population-based data, this work encompasses mammographic findings and demographic characteristics to the model. Even though this work doesn't fit general statement regarding the search criteria, I decided to include it in this case study because no other search result met our needs. The major intense in developing decision-making tools that can discriminate among benign and malignant findings in breast cancer is commented by the authors. They also mention that when developing prediction models, risk stratification is of major interest. According to their knowledge, existing studies based on the use of computer models, have also utilized specific ML techniques, such as ANNs, in order to assess the risk of breast cancer patients. In their work, ANNs are employed in order to develop a prediction model that could classify malignant mammographic findings from benign. They built their model with a large number of hidden layers which generalizes better than networks with small number of hidden nodes. Regarding the collected data in this study, mammographic findings as well as demographic risks factors and tumor characteristics were considered. All of the mammographic records were reviewed by radiologists and the reading information was obtained. This dataset was then fed as input to the ANN model. Its performance was estimated by means of ten-fold cross validation. Additionally, in order to prevent the case of overfitting the authors used the ES approach. This procedure, generally, controls the network error during training and stops it if overfitting occurs. The calculated AUC of their model was 0.965 following training and testing by means of ten-fold cross validation. The authors claimed that their model can accurately estimate the risk assessment of breast cancer patients by integrating a large data sample. They also declared that their model is unique among others if we consider that the most important factors they used to train the ANN model are the mammography findings with tumor registry outcomes. One very interesting characteristic in this study is the calculation of two main components of accuracy, namely discrimination and calibration. Discrimination is a metric that someone calculates in order to separate benign abnormalities from malignant ones, while calibration is a measurement used when a risk prediction model aims to stratify patients into high or low risk categories.

Based on my survey, I here present the most relevant and recent publications that proposed the use of ML techniques for cancer recurrence prediction. A work which studies the recurrence prediction of oral squamous cell carcinoma (OSCC). They suggested a multiparametric Decision Support System in order to analyze the basis of OSCC evolvement after total remission of cancer patients. They exploited heterogeneous sources of data (clinical, imaging and genomic) in order to predict a possible relapse of OSCC and thus a subsequent recurrence. A total number of 86 patients were considered in this study, 13 of which have been identified with a relapse while the remaining was disease free. A specific feature selection procedure was followed with the employment of two feature selection algorithms, namely CFS and wrapper algorithm. As a result, any bias could be avoided when selecting the most informative features of their reference heterogeneous dataset. Then the selected important variables could be used as input vectors to specific classifiers. Before the employment of the feature selection techniques the total number of the clinical, imaging and genomic features was in each category. Subsequently, after the employment of the CFS algorithm the total number of clinical, imaging and genomic data used in each classifier respectively. More specifically, among the clinical variables the most informative ones, for each [classification algorithm](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/classification-algorithm), were the smoker, tumor thickness and p53 stain. Concerning the imaging and the genomic features, after the utilization of the CFS algorithm, the most important were the extra-tumor spreading, the number of lymph nodes and the [SOD2](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sod2), TCAM and OXCT2 genes.

The basic idea in this study is summarized in the discrimination of patients into those with a disease relapse and those without after the performance of five classification algorithms. The employed algorithms include the BNs, ANNs, SVMs, DTs and RF classifiers. After the performance of each ML method an evaluation technique, namely ten-fold cross-validation, was employed for evaluation purposes. Additionally, accuracy, sensitivity and specificity were also calculated for comparison reasons among the employed classification schemes. The analysis of ROC curve was considered by the authors for evaluation purposes as well. Their predictive results regarding the classification schemes employed were obtained based on the classification of data without performing feature selection and on the classification of data after employing a feature selection algorithm. Regarding their outputs the authors claimed that the BN classifier without applying any feature selection scheme performed better in the discrimination with directly input of the clinical and imaging features (78.6% and 82.8% accuracy, respectively). In a similar manner, genomic-based classification results revealed that the best performing classifier was the BN in conjunction with the CFS algorithm (91.7% accuracy). In the final stage of their study, the authors combined the more accurate individual predictors (i.e. BN and BN coupled with the CFS) in order to yield a consensus decision for discrimination between patients with and without an OSCC relapse. A comparison of this approach to other studies in the literature revealed that this proposal yields robust results than other methodologies. The proposed study illustrated in an explanatory way how the integration of heterogeneous sources of data, by means of ML classifiers, can produce accurate results regarding the prediction of cancer recurrence. Furthermore, the authors used more than one classification technique in order to obtain robust results. It is clear that when you estimate the performance of a classifier predictor among others, then you are able to find the most optimal tool. However, we should highlight an important aspect of this work regarding the small sample size. Only 86 patients were considered with their clinical, imaging and genomic features. Although their classification results were very promising, we should consider that a relatively small sample size compared to data dimensionality can lead to misclassification and biased predictors. Another interesting article published in the same year with proposed an SVM-based model for the prediction of breast cancer recurrence, called BCRSVM . The authors support the idea that the classification of cancer patients into high-risk or low-risk groups allows experts to adjust a better treatment and follow-up planning. Their study is based on the development of a predictive model regarding the breast cancer recurrence within five years after surgery. SVM, ANN as well as Cox-proportional hazard regression were employed for producing the models and find the optimal one. The authors claimed that after comparing the three models based on their resulted accuracies, they found that the BCRSVM model outperformed the other two. From the initial set of 193 available variables in their dataset, only 14 features were selected based on their clinical knowledge. These data refer to clinical, epidemiological and pathological variables of 733 patients considered out of 1.541. In the final stage of the feature selection, Kaplan–Meier analysis and Cox regression were applied which resulted in 7 variables as most informative. These features were then entered as input to the SVM and ANN classifiers as well as to the Cox regression statistical model. In order to evaluate the performance of the models, the authors employed the hold-out method, which splits the data sample into two sub-sets, namely training and testing set. As in most studies in the literature, accuracy, sensitivity and specificity were calculated for a reliable estimation of the models. Based on these metrics, the authors claimed that BCRSVM outperformed the ANN and Cox regression models with accuracy 84.6%, 81.4% and 72.6%, respectively. Comparison among the performance of other previously established recurrence prediction models revealed that BCRSVM has superior performance. It should be noted that this study estimated also the importance of prognostic factors by means of normalized mutual information index (NMI). Based on these calculations for each of the three predictive models, they suggest that the most significant factor regarding the prediction of breast cancer recurrence was the local invasion of tumor. However, if someone reviews this work would certainly mention some major limitations. As the authors noted, the exclusion of a large number of patients (n = 808) due to the lack of clinical data in the research registry, influenced the performance of their models. Furthermore, the fact that the authors used only their clinical knowledge to select 14 out of 193 variables may have resulted in significant bias, thus giving no robust results. Apart from this limitation, the authors could also improve the performance of their proposed model, namely BCRSVM, by validating it with external datasets from other sources. Among the initial list of publications resulted from our literature survey, we noticed a growing trend the last years regarding the prediction of cancer disease by means of SSL learning. So, we believed it would be of interest to present the most recent study that makes use of this type of ML techniques for the analysis of breast cancer recurrence. The proposed algorithm is based on the use of SSL for the construction of a graph model while it integrates gene expression data with gene network information in order to predict cancer recurrence. Based on biological knowledge, the authors selected gene pairs that indicate strong biological interactions. The sub-gene network identified by the proposed method is constituted of the [BRCA1](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/brca1), CCND1, [STAT1](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/stat1) and CCNB1 genes. Their methodology is divided in three sections including: (i) the determination of gene pairs for building the graph model with only labeled samples, (ii) the development of sample graphs based on informative genes and (iii) the regularization of the graph resulting in finding the labels of unlabeled samples. These samples were classified into three groups: (i) recurrence, (ii) non-recurrence and (iii) unlabeled samples and referred to cancer types like breast and colon cancer. Additionally, they downloaded from the I2D database a sample of human PPIs composed of 194.988 known, experimental and predicted interactions. After removing the duplicated PPIs and the interactions that do not contain proteins mapped to a gene they resulted in an amount of 108,544 interactions. Based on the results of this study, the authors showed that the gene networks derived from the SSL learning method include many important genes related to cancer recurrence. They also claimed that their approach outperforms other existing methods in the case of breast cancer recurrence prediction. The estimated performance of the proposed method compared to other known methods that make use of PPIs for the identification of informative genes showed an accuracy of 80.7% and 76.7% in the breast and colon cancer samples, respectively. Ten-fold cross validation was used for estimating the experimental results. Although this type of ML methods differs considerably from these of supervised and unsupervised learning on the algorithms that they employ, it is clear that it provides more advantages relevant to the collection of datasets and their sizes. Unlabeled data are cheap and can be easier extracted. On the contrary, labeled samples may require experts and special devices in order to be collected. This study reveals that SSL can be an alternative to supervised approaches which usually suffers from small labeled samples.

In a predictive model is developed for the evaluation of survival in women that have been diagnosed with breast cancer, while they addressed the importance of robustness under the model's parameter variation. They compared three classification models namely SVM, ANN and SSL based on the SEER cancer database. The dataset is composed of 162,500 records with 16 key features. A class variable was also considered, namely survivability, referring to patients that had not survived and those that had survived. Among the most informative features are (i) the tumor size, (ii) the number of nodes and (iii) the age at diagnosis. By comparing the best performance of each of the three models they found that the calculated accuracy for ANN, SVM and SSL was 65%, 51% and 71% respectively. Five-fold cross validation was used for evaluating the performance of the predictive models. Concerning those findings the authors proposed the SSL model as a good candidate for survival analysis by the clinical experts. We should note that no preprocessing steps were mentioned by the authors regarding the collection of the most informative features. They proceeded with the entire SEER datasets and the box-whisper-plot was used for estimating the performance variation across 25 combinations of model parameters. A small box area of a specific model indicates more robustness and stability under parameter combination. The small boxes of the SSL model revealed its better accuracy than the other models. A relevant study was published the next year which attempts to assess the survival prediction of non-small cell lung cancer (NSCLC patients through the use of ANNs. Their dataset consists of NSCLC patients' gene expression raw data and clinical data obtained from the NCLC array database . After the preprocessing steps in their approach, the authors selected the most informative survival-associated gene signatures; LCK and ERBB2 genes, which were then used for training the ANN network. Four clinical variables, namely sex, age, T stage and N stage were also considered as input variables in the ANN model. They also performed several types of ANN architectures in order to find the optimal one for the prediction of cancer survival. An overall accuracy of 83% was provided regarding the predictive performance of the classification scheme. Furthermore, their results revealed that all patients were classified in different groups regarding their treatment protocol while 50% of them had not survived. The evaluation of the model outcomes was done based on the Kaplan–Meier survival analysis. They estimated the survival of patients for the training set, the test set and the validation set with p-value < 0.00001, while they showed that the patients in the high-risk group exhibited a lower median overall survival in comparison to low-risk patients. Compared to other studies in the literature relevant to NSCLC survival prediction, this work provided more stable results. However, existing limitations of the current article are related to the fact that the impact of other variables related to death (such as blood clots) is not considered, which may have led to misclassification results. Furthermore, the authors claim that their model could not be applied to other cancer types except NSCLC. This assumption is considered as a major limitation in studies that the predictive models may not generalize to different cancer types.

In the present review, the most recent works relevant to cancer prediction/prognosis by means of ML techniques are presented. After a brief description of the ML branch and the concepts of the data preprocessing methods, the feature selection techniques and the classification algorithms being used, we outlined three specific case studies regarding the prediction of cancer susceptibility, cancer recurrence and cancer survival based on popular ML tools. Obviously, there is a large amount of ML studies published in the last decade that provide accurate results concerning the specific predictive cancer outcomes.

**CHAPTER THREE: RESEARCH METHODOLOGY.**

**3.0: Introduction**.

This chapter describes the methods in which research is conducted, data collection techniques, tools available for making the research into reality and some of the essential ways of achieving the objectives of the research.

An incremental approach is taken which follows a set of steps and procedure when conducting research.

1.Identifying the problem. This is not only in this case but oftentimes in all cases the first step in conducting a research. Identifying the problem is what gives motivation and the urge to research about the problem. The major problem towards this research was the increased use of websites to phish sensitive data from web users and launching attacks.

2. Reviewing the literature: Review of literature is done by basing on previous researches done about phishing detection and use of machine learning to identify these phishers and other artificial intelligence techniques, how they achieved desired objectives and disadvantages. This brings the need to dive into the new ways of doing the research to improve and explore more ways of detecting the phishing websites.

3. Defining the problem through procedures: Having a deeper outlook into the problem leads to the study of how different mechanisms of solving the problem work. The workings of the different mechanisms of solving problems bring about the understanding of how the problem is really solved. This includes coming up with more procedural approaches to solve the identified problem.

4. Analysis of the collected data: Data collected in the process of problem definition is then analyzed and bring in the comparison aspect of the research method. Understanding the problem and the solution is what brings about a conclusion and the way of solving problem.

**3.1: Research Design.**

This describes the research structure employed in carrying out the research. This study makes use of both qualitative and quantitative research strategy. There will be both numerical and descriptive or conceptual data to deliver the objectives of the research study. Since cancer is a public health issue, effective countermeasures are built for different aspects in terms of education, legal supervision, and technical approaches. This survey focuses on technical strategies for detecting cancer. The methodologies of detecting cancer are developed, which are divided into three categories:

* Imaging tests.
* Biopsy.
* Machine learning methods.

Imaging testing include X-Ray, Computerized tomography Scan, Magnetic resonance imaging etc.

**3.2: Research Tool and Procedure.**

This basically describes data and information collection and design of the proposed methodology.

**3.2.1: Data Collection and analysis.**

The methods used to obtain information about this research study is an essential tool in the research since it helps give insight on how the data is useful and what are the exactly needs for the achievement of the study.

Data collection is the process of gathering and measuring information related to the study that helps in answering the research questions. The methods are varied in terms of time, cost of money or other resources at disposal of researcher (Orodho, 2008). The methods include questionnaires, personal interviews that are face to face or through the telephone or through direct observations.

To achieve the study, extensive research on existing literature articles from the internet has been a greater tool in data collection of the relevant information used in the research study, observations and interviews also have been used to gain information about cancer.

**3.2.1.1: Study.**

According to (Mohammed Baqer M. Kamel1, Loay E. George, 2000) research article, the internet and machine learning has been one of the most important study fields as it is related directly to the human's life since most people are exposed and conversant to the use of internet.

## This was based on already published books and research publications that already have the summarization of how researchers have been able to predict cancer using machine learning and other methodologies.

**3.2.1.2: Interviews and observations.**

This has been a greater method also in getting insight of the research and experiences from some of the students who may have encountered cancer patients be it relatives or friends or have general knowledge on cancer through education.

Some fellow students shared their experiences o. Some of the questions that were asked include the following;

1. Who can get cancer?
2. How does cancer starts?
3. Is cancer genetic?
4. Do you have any experience on cancer maybe with anyone around you?
5. Have you heard of machine learning and if so, what did you hear about it?

**3.2.2: Dataset.**

The dataset used in this paper was downloaded from UCI Machine Learning Repository Center for Machine Learning and Intelligent System. It contains the types of cancer and Lung cancer, Breast cancer, Cervical cancer, Prostate cancer, among others.

1. Lung cancer- Blood samples.
2. Breast cancer-Inversion of Tumor.
3. Cervical cancer- Tumor.
4. Oral cancer- Stain, Tumor spreading.

**3.2.3: Classifier.**

This work used the above data set to compare the performance of four classifiers. Specifically, we used the decision tree, SVM and BN.

**1.Decision Tree.**

Decision trees follow a tree structured classification scheme where the node represents the input variables and leaves corresponds to decision outcome.

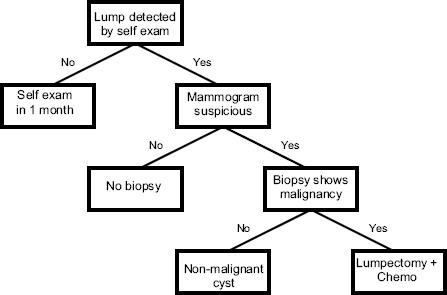


Fig.1. An illustration a simple DT showing a tree structure that can be used in breast cancer prognosis.

**2. SVM**

SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained for SVM.

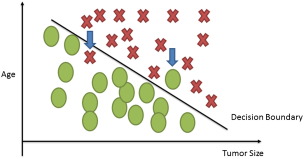


Fig.2.illustrates how an SVM might work in order to classify tumors among benign and malignant based on their size and patients' age.

**3. BN**

BN classifiers produce probability estimations rather than predictions. As their name reveals, they are used to represent knowledge coupled with probabilistic dependencies among the variables of interest via a directed acyclic graph.

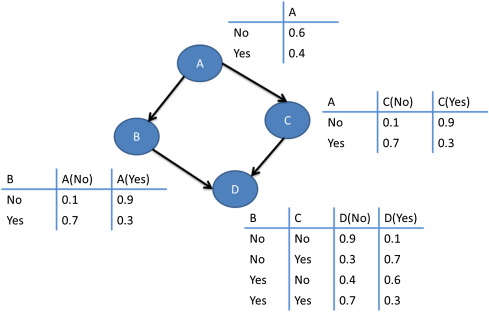


Fig. 3. An illustration of a BN. Nodes (A–D) represent a set of random variables across with their conditional probabilities which are calculated in each table.

List of events:

Lump detected by self exam- A

Self examination- B

Mammogram suspicious- C

Size of the Tumor- D

**CHAPTER FOUR: RESULTS AND DISCUSSION.**

**4.1 Introduction**

This chapter reports major findings of the study which was tested.

**4.2 Findings and Data Analysis**

We have studied some of the recognizable ML techniques and integrated data from heterogeneous sources in order to predict the desirable outcome. Table 4.2.1  depict some of the findings acquired during study. Cancer type, ML method, number of patients, type of data as well as the overall accuracy achieved by each proposed method are presented.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML Algorithm | Cancer Type | Number of Patients | Type of Data | Accuracy | Important Features |
| BN | Oral | 9 | Clinical imaging | 100% | Smoke, stain, tumor spreading |
| SVM | Breast | 15 | Clinical | 89% | Tumor |
| SVM | Cervical | 7 | Clinical | 91% | Cell type target |
| DT | Breast | 15 | Clinical | 93% | Age at diagnosis, tumor size, family history |

Table 4.2.1: Results and findings

Top of Form

Bottom of FormIn the above table, we note the accuracy presented in each and every ML algorithms used basing on the type cancer tested. A total number of 50 patients were considered during this study, 31 who were identified with the disease while 19 were cancer free. Before employing feature selection techniques, the number of clinical imaging was collected which was 41 in total.

The basic idea in this study is summarized in the discrimination of patients into those with a disease and those without after the performance of five classification algorithms. The employed algorithms include the BNs, SVM and DTs classifiers. Additionally, were also calculated for comparison reasons among the employed classification schemes. The analysis of ROC curve was considered for evaluation purposes as well. Their predictive results regarding the classification schemes employed were obtained based on the classification of data without performing feature selection and on the classification of data after employing a feature selection algorithm. Regarding their outputs we claimed that the BN classifier without applying any feature selection scheme performed better in the discrimination with directly input of the clinical and imaging features 100% accuracy, DT with 93% accuracy and SVM with 89% and 91% accuracy.

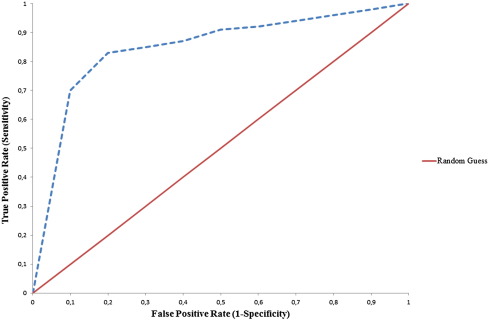


Fig. 4. An indicative ROC curve of two classifiers: (a) Random Guess classifier (red curve) and (b) A classifier providing more robust predictions (blue dotted curve)

# **CHAPTER 5: CONCLUSION, RECOMMENDATIONS, CHALLENGES AND FUTURE WORK.**

**5.1: Conclusion.**

In this review, we discussed the concepts of ML while we outlined their application in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using [supervised ML](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/supervised-machine-learning) methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.

**5.2: Recommendation**.

**5.3: Challenges and Future Work**

The major challenges in achieving this research paper include the following

* **Small Amount of Data Samples**

A basic requirement when using classification schemes for modelling disease is the size of training datasets that needs to be sufficiently large. A relative large dataset allows sufficient partitioning into training and testing sets, thus leading to reasonable validation of the estimators. A small training samples, compared to data dimensionally, can result to misclassifications while the estimators may produce unstable and biased models

* **Dataset Quality**

The dataset quality as well as careful feature features selection schemes are of great importance for effective ML and for accurate cancer predictions. Choosing the most informative feature subset for training a model, by means of feature selection methods could result in robust models. Moreover, feature sets that consist of histological or pathological assessment are characterized by reproducible value, Due to lack of statistical entities when dealing with clinical variables it is important for ML technique to be adjusted to different feature sets over time.

* **Small and Dependent Feature**

Most of the work presented here performed a validation test for estimating the performance of their learning algorithms. We employed well known evaluation techniques that split initial the initial datasets into subsets. In order to obtain accurate results for their predictive models, we should select large and independent features that could result in better validation. External and internal validation would enable extraction of more accurate and reliable predictions while it minimizes any bias.

For future works, new methods should be studied in order to overcome the limitations stated above, A better statistical analysis of the heterogenous datasets used would provide more accurate results and would give reasoning to disease outcomes. Further research is required based on construction of more public databases that would collect valid cancer datasets of all patients diagnosed with the disease. Their exploitation by researchers would facilitate their modeling studies resulting in more valid results and integrated clinical decision making.

**REFERENCES**

**1.** S. Shandilya and C. Chandankhede, "Survey on recent cancer classification systems for cancer diagnosis", *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, pp. 2590-2594, Mar. 2017.

2.Chakraborty A, Kar AK (2017) Swarm intelligence: a review of algorithms. In: Nature-inspired computing and optimization. Springer, pp 475–494

3.Yamada M, Lian W, Goyal A, Chen J, Wimalawarne K, Khan SA, Chang Y

(2017) Convex factorization machine for toxicogenomics prediction. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1215–1224

4.Wang L, Li X, Zhang L, Gao Q (2017) Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer 17(1):513–524

5.Haixiang G, Yijing Li, Jennifer Shang Gu, Mingyun HY, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239

6.Dagogo-Jack I, Shaw AT (2018) Tumor heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15(2):81–94

7.Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A (2019) Feature selection methods and genomic big data: a systematic review. J Big Data 6(79):1–24

8.Sharif MI, Li JP, Naz J, Rashid I (2020) A comprehensive review on multi-organs tumor detection based on machine learning. Pattern Recognit Lett 131:30–37

9.Dwivedi AK (2018) Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Comput Appl 29(12):1545–1554

10. UCI ML Repository.