REGRESSION MODEL FOR PREDICTING BREAST CANCER PATIENTS USING INTEGRATED GENOMIC DATA IN KENYA: A CASE OF KENYATTA NATIONAL HOSPITAL

SUBMITTED BY:

DOREEN BUNDI

20/00671

SUPERVISOR:

DR. STEPHEN NJENGA

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTERS DEGREE IN DATA ANALYTICS AT KCA UNIVERSITY

OCTOBER 2021

DECLARATION

This project is my own work and it has not in part or fully been submitted or presented for award of degree or any other academic work

Signature……  …….Date ……11/06/2021……..

DOREEN BUNDI

20/00671

This project has been submitted for examination with my approval as the appointed university supervisor.

DR. STEPHEN NJENGA

Signature: …… …… Date: ……11/06/2021……

ABSTRACT

Cancer has been characterized as a heterogeneous disease which has caused havoc worldwide with the increasing deaths related to cancer. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. The main objective of the study was to develop a regression model for predicting breast cancer patients using integrated genomic data. It was facilitated by the objectives that sought to review the literature on factors to predict breast cancer patients using integrated genomic data, develop a regression model for predicting breast cancer patients using integrated genomic data and test and validate the regression model for predicting breast cancer patients using integrated genomic data. Data was obtained online through openML site. Information will be abstracted from the data obtained was used for assessing breast cancer patients. The researcher utilized the Kenyatta National Hospital dataset that includes 44,000 cancer patients. This formed the target population used in the study. The population was narrowed down to 1,172 new cases between January of 2017 and June 2019. Analysis was conducted by reviewing the literature, assessing the details and testing and validating the model for predicting cancer patients using integrated genomic data machine learning model will be applied. Inferential data analysis was used in reviewing the literature. In this case, the data was summarized into points in a constructive manner. The analysis was vital in forming the basis of quantitative data analysis. Regression analysis was used in the identification of supervised learning models and their influence on the topic. Additionally, regression analysis was employed as a predictive modeling technique that assesses the affiliation between the variables. The research findings further established that factors influencing breast cancer prognosis, screening appropriate predictors as independent variables are an important step in model construction. In this case, the demographic risk factors are important in the creation of BC risk prediction model. Additionally, it was found that the genetic variants, combinations of demographic risk factors yielded a higher risk prediction accuracy than the individual demographic risk factors. Age, disease stage, grade, tumor size, race, marital status, number of nodes, histology, number of positive nodes and primary site code have been entered into many predictive models as predictors, given that these factors represent key risk factors for onset and survival in breast cancer. The researcher proposed an ML approach to efficiently combine genetic variants with BC risk factors related to both familial history and oestrogen metabolism and to search for optimal interactions among them. According to the research, the choice of the most appropriate algorithm depends on many parameters including the types of data collected, the size of the data samples, the time limitations as well as the type of prediction outcomes. Therefore, it was recommended that the future of cancer modeling new methods should be studied for overcoming the limitations.

ACKNOWLEDGEMENT

I acknowledge the Almighty God for His gift of life. This research project was a result of support from several sources; I would like to acknowledge my supervisor for continuous encouragement, patience and keen guidance through the project and continued interest in my study. Also, much appreciation to KCA University and the staff for the conducive learning environment during my study period and as a source of knowledge and making this research work successful.

ACRONYMS AND ABBREVIATION

ML	Machine Learning
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
MiRNA	Micro Ribonucleic acid
SNPs	Single nucleotide polymorphisms
TCGA	The Cancer Genome Atlas
GEO	Gene Expression Omnibus
WBC	White blood cells
ANN	Artificial neural network
CNN	Convolution neural network

GLOSSARY

BRCA1/2 Denotes the genes commonly affected in hereditary breast and ovarian cancer are the breast cancer 1 (BRCA1) and breast cancer 2 (BRCA2) genes.

NF1 Denotes as a genetic condition that causes tumors to grow along your nerves. The tumors are usually non-cancerous (benign) but may cause a range of symptoms.

RB1 The RB1 gene provides instructions for making a protein called pRB. This protein acts as a tumor suppressor, which means that it regulates cell growth and keeps cells from dividing too fast or in an uncontrolled way.

CDK12 Cyclin-dependent kinase 12 (CDK12) is an important transcription associated CDK. It shows versatile roles in regulating gene transcription, RNA splicing, translation, DNA damage response (DDR), cell cycle progression and cell proliferation.

TP53 The TP53 gene provides instructions for making a protein called tumor protein p53 (or p53). This protein acts as a tumor suppressor, which means that it regulates cell division by keeping cells from growing and dividing (proliferating) too fast or in an uncontrolled way.

FOXM1 Fork head box protein M1 is a transcription factor required for a wide spectrum of essential biological functions, including DNA damage repair, cell proliferation, cell cycle progression, cell renewal, cell differentiation and tissue homeostasis.

TABLE OF CONTENTS

DECLARATION ii
ABSTRACT iii
ACKNOWLEDGEMENT iv
GLOSSARY vi
TABLE OF CONTENTS vii
LIST OF FIGURES ix
List of Tables x
CHAPTER ONE ..... - 1 -
INTRODUCTION ..... - 1 -
1.1 Background of the study ..... - 1 -
1.2 Statement of the Problem ..... - 7 -
1.3 Main Objective ..... - 8 -
1.3.1 Specific Objectives ..... - 8 -
1.3.2 Research questions ..... - 8 -
1.4 Significance of the Study ..... - 8 -
1.6 Scope of the study ..... - 10 -
CHAPTER TWO ..... - 11 -
LITERATURE REVIEW ..... - 11 -
2.1 Introduction ..... - 11 -
2.2 Theoretical review ..... - 11 -
2.2.1 Biological theory ..... - 11 -
2.2.2 Machine learning theory ..... - 12 -
2.3.1 Cancer ..... - 15 -
2.3.2 Machine Learning ..... - 16 -
2.3.3 Deep learning architectures ..... - 16 -
2.3.4 Machine learning models ..... - 17 -
2.3.5 Existing cancer prediction models ..... - 20 -
2.4. Type of Cancer ..... - 24 -
2.5 Gender ..... - 25 -
2.6 Tumor Stage ..... - 25 -
2.7 Age ..... - 26 -
2.8 Hereditary Factor ..... - 27 -
2.9 Conceptual Framework ..... - 28 -
2.10 Operationalization of Variables ..... - 29 -
2.11 Summary ..... - 30 -
CHAPTER THREE ..... - 31 -
METHODOLOGY ..... - 31 -
3.1. Introduction ..... - 31 -
3.2 Research design ..... - 31 -
3.3. Target population ..... - 32 -
3.4. Research Instrument ..... - 32 -
3.5. Validity and Reliability of the instrument ..... - 32 -
3.6. Data collection procedure ..... - 33 -
3.7. Data processing and analysis ..... - 33 -
3.8 Model Development Process ..... - 34 -
CHAPTER FOUR ..... - 38 -
RESEARCH FINDINGS AND DISCUSSION ..... - 38 -
4.1 Introduction ..... - 38 -

4.2 Demographic Information ..... - 38 -
Type of Cancer and Gender ..... - 38 -
4.3 Research Findings ..... - 43 -
4.3.1 Factors to Predict Breast Cancer Patients using Genomic Data ..... - 43 -
4.3.2 Results for Regression Model for Predicting Breast Cancer Patients using Genomic Data ..... - 45 -
4.4.3 Results for Test and Validation of the Model ..... - 47 -
4.6 Discussion of Results ..... - 48 -
4.5 Summary ..... - 51 -
CHAPTER FIVE ..... - 52 -
CONCLUSIONS AND RECOMMENDATION ..... - 52 -
5.1 Introduction ..... - 52 -
Conclusion ..... - 52 -
5.3 Contributions of the Study ..... - 55 -
5.3.1 Model Comparison ..... - 56 -
5.4 Recommendations for Future Research ..... - 56 -
References ..... - 58 -
APPENDICES ..... - 62 -
Appendix 1: Research schedule ..... - 62 -
Appendix ii: Resources and Budget ..... - 63 -

LIST OF FIGURES

Figure 2.1: The Conceptual Framework ..... - 28 -

Figure 3.1 Overview of Proposed Method ..... - 37 -

Figure 4.2 Proposed BC Risk Prediction Model Architecture. ..... - 46 -

List of Tables

Table 2.1 Operational Variables ..... - 28 -

Table 3.1 Analysis of each Objective ..... - 34 -

Table 4.1: Demographic Characteristics ..... - 38 -

Table 4.2: Age ..... - 39 -

Table 4.3: Tumor Stage ..... - 40 -

Table 4.4 Distribution of BC Risk Factors in Familial History ..... - 41 -

Table 4.5: Distribution of Malignant Lesions ..... - 43 -

Table 4.6: Distribution of Benign Breast Cancer ..... - 44 -

Table 4.7: Distribution of Male Breast Pathology ..... - 44 -

Table 4.8: Mode of Initial Definitive Diagnosis ..... - 45 -

Table 4.9 Primary Treatment Offered Initially ..... - 45 -

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients (Das et. al. 2019). The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods (Davi, & Acioli-Santos, 2019). Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance (Tresp, & Yu, 2016). Even though it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice.

Breast cancer affects many people at the present time. The factors that cause this disease are many and cannot be easily determined. Additionally, the diagnosis process which determines whether the cancer is benign or malignant also requires a great deal of effort from a doctors and physicians. When several tests are involved in the diagnosis of breast cancer, such as clump thickness, uniformity of cell size, uniformity of cell shape...etc, the ultimate result may be difficult to obtain, even for medical experts. This has given a rise in the last few years to the use of machine learning and Artificial Intelligence in general as diagnostic tools.

Epidemiological studies reveal wide disparities in the frequency and distribution of breast ailments across the world (Otieno, 2008). When local breast disease distribution patterns are known, generalizations pertaining to diagnosis and management can be made with a reasonable degree of certainty. In addition, resource allocation and planning can be better managed. This is particularly so in resource poor countries where a large population of individuals may not afford all the forms of diagnostic modalities available.

Breast diseases afflict women more than men, the prevalence rate in males ranging from 0 to 5.8% in most series. Majority of male breast afflictions are benign with gynaecomastia occupying the top slot. Among females, the distribution of pathology varies widely depending on age and geographical location. Benign lesions predominate at all ages accounting for 48.9% to 57% with a mean age of occurrence being 28.5 years. Benign lesion prevalence rates can peak 99% in those younger than 30 years. Fibroadenoma is the most common lesion at prevalence rates between 34.7 to 67% of all breast lesions with a peak mean age incidence of 16-25 years. Malignancy or inflammatory lesions come second in frequency to fibroadenoma. Benign diseases thus, constitute the major work load in any breast clinic, although some studies have shown that malignant conditions predominate, yet others have found inflammatory lesions to be the most common disease entity.

Cancer is driven by changes at the cellular and molecular levels. Its development and proliferation are associated with the accumulation of mutations. Notably, quite a few of identified mutations are responsible for cellular variations leading to cancer. Most variations are neutral and benign (passenger) in nature while a small fraction of the mutations drive the cancer development process. Accumulation of genetic alteration can result in tumor development in oncogenes, tumor-suppressor genes and stability genes (Vogelstein and Kinzler 2004; Vogelstein et al., 2013; Feinberg et al. 2006).

Mutation is the primary source of genetic variation, however sexual reproduction and recombination contribute significantly to genetic variation. Genetic variation describes the mutation in the genome's DNA sequence and is responsible for the distinct traits' humans' exhibit. It is the result of subtle differences in the DNA (Carleo et al, 2019). Variation occurs in germ and somatic cells. The only variation that arises in germ cells can be inherited from one individual to another and so affect the dynamics of the population, consequently leading to evolutionary changes. New mutations occur when there are errors during DNA replication that are not corrected by DNA repair enzymes. Most somatic mutations are salient but can occasionally interfere with major cellular functions. Early somatic mutation can lead to developmental disorders whereas incessant accumulation of mutation can cause cancer. Human cells innately have several safety protocols to protect themselves against the lethal effects of mutation inducing cancers. Therefore, it is the defective genes that result in cancer proliferation (Yeang et al., 2008; Li et al., 2016).

The identification of driver mutations in cancer remains a challenging task. There are several computational strategies aimed at detecting driver genes and ranking mutations for their carcinogenicity prospect (Torkamani et al., 2009). Consequently, several driver genes are not tagged as disease-related (Brown et al., 2019). Presently, the reoccurrence of a mutation in patients remains one of the top reliable markers of mutation driver status. Nonetheless, some mutations are more likely to occur than others due to differences in background mutations rates arising from different types of DNA replication and repair systems (Brown et al., 2019). From the study of Brown et al., 2019, they showed that mutations not yet observed in a tumor had relatively low mutability, thus indicating that background mutability might limit the occurrence of mutation.

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. Various aspects regarding the prediction of cancer outcome based on gene expression signatures are discussed in (Zairis, 2018). These studies list the potential as well as the limitations of 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. This study will examine how to develop a machine learning model for predicting cancer patients using integrated genomic data.

Machine learning is a technique used to provide computers the ability to learn without being explicitly programmed. Tom M. Mitchell provided the formal definition: ”A computer program is said to learn from experience E with respect to some class of tasks T and performance P if its performance at tasks in T, as measured by P, improves with experience E”. Machine learning algorithms build mathematical models and are closely related to computational statistics, which mainly focuses on making predictions using computers (Cammarota et al, 2020).

With time, machine learning has made remarkable contributions in bioinformatics, detecting cancer, creating new drugs, analyzing traffic patterns. However, it becomes difficult when data grow complex and huge, but the rise and advancement of machine learning algorithms have made it possible to solve many problems (Carleo et al, 2019). The area of Machine Learning deals with the design of programs that can learn rules from data, adapt to

changes, and improve performance with experience. In addition to being one of the initial dreams of Computer Science, Machine Learning has become crucial as computers are expected to solve increasingly complex problems and become more integrated into our daily lives.

When applying a ML model, data samples constitute the basic components. Every sample is described with several features and every feature consists of different types of values (Zairis, 2018). 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 pre-processing 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 (Libes et. al. 2014). In addition, in order to make the raw data more suitable for further analysis, pre-processing steps should be applied that focus on the modification of the data. A number of different techniques and strategies exist, relevant to data pre-processing that focus on modifying the data for better fitting in a specific ML method.

Over the past twenty years, the types of applications of machine learning have grown more and more varied ranging from computational biology to astronomy to robotic surgery. Moreover, many of the (new and old) application areas have faced a huge increase in the volume of available data of various kinds. In order to better use all the available data a number of powerful new learning approaches have been proposed. These approaches have been intensely explored in the machine learning community, with many heuristics and specific algorithms, as well as various successful experimental results reported (Das et. al. 2019). Unfortunately, however, the standard theoretical models do not capture the key issues involved in these learning techniques, and it has become clear that for developing robust, versatile, and general algorithms in these settings a general fundamental understanding is necessary. In this thesis we develop such theoretical foundations as well as new and general algorithms for these emerging machines learning paradigms, including Semi-Supervised, Active, and Similarity-based Learning. In addition, the novel insights we develop here allow us to also revisit the classic problem of Clustering which has not been satisfactorily captured by existing models. This dissertation also brings forward new connections between Machine

Learning and Algorithmic Game Theory, an emerging area at the intersection of Computer Science and Economics.

Machine learning has been applied to many areas in health care, including imaging diagnosis, digital pathology, prediction of hospital admission, drug design, classification of cancer and stromal cells, doctor assistance. Cancer prognosis is to estimate the fate of cancer, probabilities of cancer recurrence and progression, and to provide survival estimation to the patients. The improvement of biomedical translational research and the application of advanced statistical analysis and machine learning methods are the driving forces to improve cancer prognosis prediction (Du & Swamy, 2019). Recent years, there is a significant increase of computational power and rapid advancement in the technology of artificial intelligence, particularly in machine learning. In addition, the cost reduction in large scale next-generation sequencing, and the availability of such data through open-source databases (e.g., TCGA and GEO databases) offers opportunities to possibly build more powerful and accurate models to predict cancer prognosis more accurately. The application of machine learning in cancer prognosis has been shown to be equivalent or better than current approaches, such as Cox-PH (Hajiloo, & Damaraju, 2013).

Machine learning methods can be used in the design of auctions and other pricing mechanisms with guarantees on their performance. Adaptive machine learning algorithms can be viewed as a model for how individuals can or should adjust to changing environments (Cammarota et al, 2020). Moreover, the development of especially fast-adapting algorithms sheds light on how approximate equilibrium states might quickly be reached in a system, even when each individual has a large number of different possible choices. In the other direction, economic issues arise in Machine Learning when not only is the computer algorithm adapting to its environment, but it also is affecting its environment and the behavior of other individuals in it as well.

Among these techniques some of the most important approaches include dimensionality reduction feature selection and feature extraction (Kundrod et. al. 2019). Machine learning algorithms are often categorized as supervised or unsupervised. Supervised algorithms require both input and desired output, in addition to furnishing feedback about the accuracy of predictions during algorithm training (Libes et. al. 2014). Once training is complete, the algorithm will apply what was learned to new data. Unsupervised algorithms do not need to be trained with desired outcome data. Since the success of a learning algorithm

depends on the data used, machine learning is inherently related to data analysis and statistics (Das et. al. 2019). More generally, machine learning techniques are data-driven methods combining fundamental concepts in computer science.

In order to more completely understand complex biological phenomena, such as many human diseases or quantitative traits in animals/plants, massive amounts and multiple types of 'big' data are generated from complicated studies. In the not-so-distant past, data generation was the bottleneck, now it is data mining, or extracting useful biological insights from large, complicated datasets. In the past decade, technological advances in data generation have advanced studies of complex biological phenomena (Cammarota et al, 2020). In particular, next generation sequencing (NGS) technologies have allowed researchers to screen changes at varying biological scales, such as genome-wide genetic variation, gene expression and small RNA abundance, epigenetic modifications, protein binding motifs, and chromosome conformation in a high-throughput and cost-efficient manner. The explosion of data, especially omics data, challenges the long-standing methodologies for data analysis with ideas from statistics, probability and optimization

The boundary between machine learning and statistics is fuzzy. Some methods are common to both domains and either can be used for prediction and inference. However, machine learning and statistics have different foci, prediction or inference. In general, classic statistical methods rely on assumptions about the data-generating systems. Statistics can provide explicit inferences through fitting a specified probability model when enough data are collected from well-designed studies (Das et. al. 2019). Machine learning is concerned with the question of creation and application of algorithms that improve with experience. Many machine learning methods can derive models for pattern recognition, classification, and prediction from existing data and do not rely on stringent assumptions about the data-generating systems, which makes them more effective in some complicated applications, as further described below, but less effective in producing explicit models with biological significance, in some cases.

Comprehensive breast cancer risk prediction models enable identifying and targeting women at high-risk, while reducing interventions in those at low-risk. Breast cancer risk prediction models used in clinical practice have low discriminatory accuracy (0.53–0.64). Machine learning (ML) offers an alternative approach to standard prediction modeling that may address current limitations and improve accuracy of those tools.

1.2 Statement of the Problem

Breast cancer (BC) is the most common cancer affecting women worldwide and the most frequent cause of cancer death in women (Otieno, 2008). Recent advances in treatment strategies have improved BC-related mortality and morbidity; however, almost 30% of BC patients show recurrence in the follow-up (Otieno, 2008). Therefore, to improve BC outcomes, it is necessary to focus on research such as improving screening methods for early detection of recurrence according to risk stratification, identifying new biomarkers, and developing new innovative treatment strategies.

There is an urgent unmet need to identify innovative methods to determine the prognosis of individual patients (Zhu, Xie, Han & Guo, 2020). Traditionally, clinicopathologic characteristics such as tumor size, axillary nodal status, histologic and nuclear grade, hormone receptors [estrogen receptor (ER) and progesterone receptor (PR)], and human epidermal growth factor receptor 2(HER2) status have been used to identify risk groups and to predict patient prognoses. In addition to clinicopathologic characteristics, multigene signature panels offer an additional benefit in predicting patient prognoses.

With the surge of medical data as well as the rapid development of information technology and artificial intelligence, the application of big data analysis technology in the construction of survival prediction model has become a current research hotspot. Traditional prediction models based on prior hypothesized knowledge often consider the relationship between dependent variables; in contrast, ML has the potential of learning data models automatically, does not require any implicit assumptions and is able to handle interdependence and nonlinear relationships between variables (Obermeyer & Emanuel, 2016). It has natural strengths in dealing with the very large number of complex higher-order interactions of medical data. Therefore, ML tools have a high potential for application in routine medical practice as leading tools in health informatics.

Existing ML models such as the Nottingham prognostic Index, The ANNs model and BOADICEA are subject to significant challenges. Adjuvant online is another computer-based prognostic model (developed in 2001) to estimate 10-year survival and recurrence for BC patient based on six variables (age at diagnosis, comorbidity, estrogen receptor (ER), tumor size, tumor grade and lymph node status) (Tsai *et al*., 2021; Babain *et al*., 2000). The models do not take into account the treatment modalities such as genomic and non-genomic factors

that are observed to be one of the strongest prognostic factors. Therefore, this study developed a regression model model of individual BC patients using the machine learning method (Tsai *et al*., 2021). This model was developed using BC-related clinicopathologic factors at the time of curative surgery and consecutive clinical factors that have been identified during the BC surveillance period.

1.3 Main Objective

The main objective of the study was to develop a regression model for predicting breast cancer patients using integrated genomic data.

1.3.1 Specific Objectives

The specific objectives of the study will be:

i) To review the literature on factors to predict breast cancer patients using integrated genomic data
ii) To develop a regression model for predicting breast cancer patients using integrated genomic data
iii) To test and validate the regression model for predicting breast cancer patients using integrated genomic data

1.3.2 Research questions

From the above objectives, the following research questions were developed:

i) What literature exists on factors to predict breast cancer patients using integrated demonic data?
ii) How can the regression model for predicting cancer patients using integrated genomic data be developed?
iii) How effective is the regression model in predicting breast cancer patients?

1.4 Significance of the Study

The proposed study will inform the government on the current situation of the perspectives on machine learning approaches to integrate genomic data and build a classifier for stratification of cancer patients in Kenya. This will enable the government to be more

proactive in implementing the technology on medical imaging to enhance medical service provision in hospitals. The study will be a source of reference material for future researchers on other related topics; it will also help other academicians who will undertake the same topic in their studies. The study will also highlight other important relationships that require further research. The findings of this research will generate information on perspectives on machine learning approaches to integrate genomic data. This will form a basis for training health workers on taking care of the patients and enhance faster delivery of medical care.1.5 Motivation of the study

Breast cancer is the most common cancer among women in 154 countries and the main cause of cancer-related death in 103 countries. In 2018, there were approximately 2.1 million new cases of breast cancer in women, accounting for 24.2% of the total cases, and the mortality rate was approximately 15.0% (Li, et al. 2021). As a result, accurately predicting the survival rate of breast cancer patients is a major issue for cancer researchers. Machine learning (ML) has attracted much attention with the hope that it could provide accurate results, but its modeling methods and prediction performance remain controversial. Hence, developing supervised learning model as a machine learning model for predicting breast cancer patients using integrated genomic data in Kenya is crucial.

While ANNs still predominate it is evident that a growing variety of alternate machine learning strategies are being used and that they are being applied to many types of cancers to predict at least three different kinds of outcomes. It is also clear that machine learning methods generally improve the performance or predictive accuracy of most prognoses, especially when compared to conventional statistical or expert-based systems (Du and Swamy, 2019). Although most studies are generally well constructed and reasonably well validated, certainly greater attention to experimental design and implementation appears to be warranted, especially with respect to the quantity and quality of biological data. Improvements in experimental design along with improved biological validation would no doubt enhance the overall quality, generality and reproducibility of many machine-based classifiers.

Although machine learning has been used primarily as an aid to cancer diagnosis and detection, a small review has been done on application of machine learning towards breast cancer prediction and prognosis. The proposed study will inform the government on the current situation of the perspectives on machine learning approaches to integrate genomic

data and build a classifier for stratification of cancer patients in Kenya (Cammarota et al, 2020). This will enable the government to be more proactive in implementing the technology on medical imaging to enhance medical service provision in hospitals.

1.6 Scope of the study

The study focused in the development and use of machine learning model in predicting cancer patients. This study will be delimited to cancer patients in Kenyatta National Hospital. Most Cancer patients in the Country, are usually referred to Kenyatta National Hospital for treatment. This is enhanced by the availability of cancer screening machines. Hence conducting the study at Kenyatta National Hospital will be viable and adequate information to further the study will be obtained. The unit of observation will be the cancer treatment center at Kenyatta National Hospital. The study will be carried out between July and October.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter discusses literature as depicted by the previous researchers based on the objectives of the study. In particular the chapter discusses the following, theoretical review, conceptual framework as well as the summary of the literature review.

2.2 Theoretical review

Several theories have given the justification on application of Machine learning model in predicting cancer patients using genomic data. This study is anchored towards the biological theory and machine learning theory. It is critical that Machine learning model is based on a solid theoretical foundation, as stipulated by Dalkir (2011), the theoretical framework provides a significant view on the variable of the study.

2.2.1 Biological theory

This theory can be classified into three types that is, those that attempt to differentiate among individuals on the basis of certain innate outward physical traits or characteristics and those that attempt to trace the source of differences to genetic or hereditary.

Biological systems are complex. Most large-scale studies focus only on one specific aspect of the biological system; for example, genome-wide association studies (GWAS) focus on genetic variants associated with measured phenotypes (Waks and Winer, 2019). However, complex biological phenomena can involve many biological aspects, both intrinsic and extrinsic and, thus, cannot be fully explained using a single data type. For this reason, the integrated analysis of different data types has been attracting more attention. Integration of different data types should, in theory, lead to a more holistic understanding of complex biological phenomena, but this is difficult due to the challenges of heterogeneous data and the implicitly noisy nature of biological data.

Machine learning has been used broadly in biological studies for prediction and discovery. With the increasing availability of more and different types of omics data, the

application of machine learning methods, especially deep learning approaches, has become more frequent. One area of opportunity for machine learning approaches is in the prediction of genomic features, particularly those that are hard to predict using current approaches such as regulatory regions (Hajiloo, & Damaraju, 2013).

Machine learning has been used to predict the sequence specificities of DNA- and RNA-binding proteins, enhancers, and other regulatory regions on data generated by one or multiple types of omics approach, such as DNase I hypersensitive sites (DNase-seq), formaldehyde-assisted isolation of regulatory elements with sequencing (FAIRE-seq), assay for transposase-accessible chromatin using sequencing (ATAC-seq), and self-transcribing active regulatory region sequencing (STARR-seq). Machine learning can be used to build models to predict regulatory elements and non-coding variant effects de novo from a DNA sequence that can then be tested/validated for their contribution to gene regulation and ultimately to observable traits/pathologies.

In addition to the prediction of regulatory regions, recently, supervised learning showed considerable potential for solving population and evolutionary genetics questions, such as the identification of regions under purifying selection or selective sweeps, as well as more complicated spatiotemporal questions Up to now, machine learning approaches have also been used to predict transcript abundance , imputation of missing SNPs and DNA methylation states ,variant calling , disease diagnosis/classification, and many different biological questions using datasets from different biological aspects such as genomes, epigones, transcriptomes, and metabolomes (Cammarota et al , 2020).

The massive and rapid advancements in both biological data generation and machine learning methodologies are promising for the analysis and discovery from complex biological data. However, there are several hurdles. Firstly, interpretation of models derived from some sophisticated machine learning approaches such as deep learning can be difficult if not impossible. In many cases, researchers are more interested in the biological meaning of the predictive model than the predictive accuracy of the model and the ‘black box’ nature of the model can inhibit interpretation.

2.2.2 Machine learning theory

Machine Learning (ML) model 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 (Hajiloo, & Damaraju, 2013). Several studies have been reported in the literature and are based on different strategies that could enable the early cancer diagnosis and prognosis.

ML offers an alternative approach to standard prediction modeling that may address current limitations and improve accuracy of breast cancer prediction tools. ML techniques developed from earlier studies of pattern recognition and computational statistical learning. They make fewer assumptions and rely on computational algorithms and models to identify complex interactions among multiple heterogeneous risk factors. This is achieved by iteratively minimizing specific objective functions of predicted and observed outcomes. ML has been used in models related to cancer prognosis and survival and produced better accuracy and reliability es- timates. However, a significant gap is presented by the fact that very few studies applied ML methods for personalized breast cancer risk prediction or compared the predictive accuracy and reliability with models commonly used in clinic practice.

A growing number of ML studies have been applied to diagnosis, disease risk prediction, recurrence prediction, and symptom prediction. Furthermore, although the number of survival predictions increases gradually, the database set, modeling process, methodological quality, performance metrics, and modeling of related candidate predictors exhibit large differences. Many state-of-the-art deep learning techniques have been applied to cancer prognosis prediction, indicating the great potential and the urgent need of utilizing multi-omics data from cancer patients to test new algorithm and improve model performance. There are seven main challenges in applying deep learning approach in cancer prognosis prediction to achieve high performance.

Machine Learning Theory, also known as Computational Learning Theory, aims to understand the fundamental principles of learning as a computational process. This field seeks to understand at a precise mathematical level what capabilities and information are fundamentally needed to learn different kinds of tasks successfully, and to understand the basic algorithmic principles involved in getting computers to learn from data and to improve performance with feedback. The goals of this theory are both to aid in the design of better

automated learning methods and to understand fundamental issues in the learning process itself (Du and Swamy, 2019).

Machine Learning Theory draws elements from both the Theory of Computation and Statistics and involves tasks such as: Creating mathematical models that capture key aspects of machine learning, in which one can analyze the inherent ease or difficulty of different types of learning problems. Second is proving guarantees for algorithms (under what conditions will they succeed, how much data and computation time is needed) and developing machine learning algorithms that provably meet desired criteria. Third is mathematically analyzing general issues, such as: When can one be confident about predictions made from limited data? (Waks and Winer, 2019)

Another highlight of Computational Learning Theory is the development of algorithms that are able to quickly learn even in the presence of large amounts of distracting information. Typically, a machine learning algorithm represents its data in terms of features: for example, a document might be represented by the set of words it contains, and an image might be represented by a list of various properties it has (Das et. al. 2019). The learning algorithm processes this information to make some prediction

There are two primary types of machine learning methods: supervised learning and unsupervised learning. Supervised learning algorithms learn the relationship between a set of input variables and a designated dependent variable or labels from training instances and can subsequently be used to predict the outcomes of new instances. Many sophisticated machine learning methods are supervised, e.g., decision tree, support vector machine, and neural network. Unsupervised learning algorithms infer patterns from data without a dependent variable or known labels. Cluster and principle component analysis are two popular unsupervised learning methods used to find patterns in high dimensionality data such as omics data. Deep learning is a subtype of machine learning originally inspired by neuroscience, essentially describing a class of large neural networks (Doria-Rose et al, 2021). Deep learning has been applied in many fields, largely driven by the massive increases in both computational power and big data. Deep learning can be both supervised and unsupervised, has revolutionized fields such as image recognition, and shows promise for applications in genomics, medicine, and healthcare.

The research will employ regression analysis as the predictive modeling technique that assesses the affiliation among two or more variables. Regression analysis will be focused on the affiliation between the dependent and independent variables. The dependent variable is assumed to the influence of the independent variables. Regression equations are a crucial part of the statistical output after you fit a model. The coefficients in the equation define the relationship between each independent variable and the dependent variable. However, you can also enter values for the independent variables. Using regression to make predictions doesn't necessarily involve predicting the future. Instead, you predict the mean of the dependent variable given specific values of the independent variable(s). The main goal is to reduce the cost functions of the model.

2.3 Empirical Review

Several researchers both locally and globally have shown interest on Machine learning model and cancer in recent studies. Cancer prognosis prediction using Machine learning algorithm has been the research trend and has attracted the interest of scholars.

2.3.1 Cancer

According to WHO, cancer is the second leading cause of death. It can be described as abnormal cells that rapidly grow in any part of the body. Cancer is a group of diseases and can appear in multiple forms and have different symptoms. There are various reasons for having cancer, such as genetic mutation and unhealthy life choices. The genetic mutation happens in the DNA amino acid sequence which changes or shift the DNA sequence structure and create mutated cells with different sequence order. There are several stages in examining possible cancer patients, such as blood work tests and physical examination (Johnson, Ben-Zion, Meng and Vernon, 2020).

Cancer mainly affects the white blood cells (WBC) and the immune system. There are five different types of white blood cells, and they are neutrophils, lymphocytes, monocytes, eosinophils, and basophils, but only the first four's level changes when the body has cancer. The WBC test works in such a way that it is performed automatically where the number of white blood cells is counted and compared with a reference table that can vary among different sites. A decreased amount of lymphocytes and Neutrophil are signs of body immune system fighting a virus, and that the body is not able to produce enough antibodies.

2.3.2 Machine Learning

Machine learning is a part of artificial intelligence, and the idea is generally defined as a software system having the information to learn from experience using a set of tasks. Three essential aspects define how machine learning functions. These aspects are tasks, experience, and performance (Waks and Winer, 2019). Tasks are datasets to train the computer to increase its performance. With time and experience, the computer system can learn and become a refined model that can prognosticate the answer to a topic that it has learned from previous attempts. There are multiple algorithms used in machine learning, but they fall into two categories, supervised learning and unsupervised learning. The supervised learning group also referred to as a method working with a set of training data. In an attempt to classify the result, the algorithm needs to work on manually entered answers. This type of working method is heavily dependent on the training data. Therefore, the set needs to be correct for the algorithm to make sense of the data.

Unsupervised learning is that the algorithm finds undetected patterns in a massive amount of data. In this type of method, it allows the computer algorithm to execute and see what the outcome patterns are going to be (Zhu, Xie, Han and Guo, 2020). For that reason, there is no clear answer that is considered right or wrong. In machine learning, there are dependent and independent variables. The independent variables are also referred to as predictor or control input; this holds the values that control the experiment. The dependent variables, otherwise known as output values, are regulated by the independent variables.

2.3.3 Deep learning architectures

Deep learning is a subsection of machine learning. It is a learning method that operates with multi-level layers and grows towards a more abstract level. The deep refers to the multiple layer in the neural network that made of nodes. Each layer in the network trained on a distinct feature based on the output from the previous layer. Deep learning is inspired by the layout of the human brain by creating the architecture base on neurons. In a human brain, there are massive amounts of neurons that are connected and create a network of communication via signals that it receives. This concept is referred to as an artificial neural network (ANN). In ANN, the algorithm creates layers that enter input values from one layer to the next, which eventually ends with an outcome result. With deep learning, humans do not interfere with the layers within a neural network and the information that is being

processed. The system algorithms are trained with data and learning procedures; therefore, it does not need to be manually handled by humans. The method gains the ability to manage higher-dimensional data. The system method has displayed a promising result in handling classification, analysis, and translations of more advanced areas

2.3.4 Machine learning models

Globally, Omar (2020) did a study on a Comparative study of cancer detection models using deep learning. A comparison study was performed by comparing two different leukemia detection methods. The methods were a genomic sequencing method, which is a binary classification model and a multi-class classification model, which was an images-processing method. The methods had different input values. However, both of them used a Convolutional neural network (CNN) as network architecture (Omar 2020). The datasets were also split using 3-way cross-validation. The evaluation methods for analyzing the results were learning curves, confusion matrix, and classification report. The results showed that the genome model had better performance and had several numbers of values that were correctly predicted with a total accuracy of 98%. This value was compared to the image processing method results that have a value of 81% total accuracy (Omar 2020). The size of the different data sets can be a cause of the different test results of the algorithms.

Maharjan (2020) studied machine learning approach for predicting cancer using gene expression. The objective of the study was to build models and classify different types of cancer. For this purpose, machine learning models such as support vector machine (SVM), random forest (RF), k-nearest neighbors (KNN) and multilayer perceptron (MLP) were implemented to classify the samples according to their labels. The machine learning models were trained on TCGA data and tested on independent dataset (GTEX) (Zhu, Xie, Han and Guo, 2020). The data representation obtained using stacked DE noising auto encoders were used to train and test the models. The models did not have very high performance; however, MLP performed better than others. The best features that were selected using SelectKBest, were also used to compare the performances. It was observed that the K-nearest neighbor classifier gave better results, with and accuracy of 85.12% while tested on independent data, and the training accuracy was 98.4%.

Another study by Dhillon, Arwinder, Kaur and amrita, (2020) on application of machine learning for prediction of lung cancer using omics data, reviewed the major

algorithms related to prior work done in the area of cancer prediction considering the different types of data taken along with the method used and their respective limitations .Comparative study of various machine learning techniques and technologies has been done over different types of data such as clinical data, omics data, image data (Dhillon, Arwinder, Kaur & amrita, 2020). Comparative study of how some algorithms work better for certain purpose has been done. The main focus was how when we consider different or heterogeneous data results have been superior. The insight and comparisons of the recent research done in this field provides useful information which can be used to predict cancer in their early stage.

Odeyemi (2020) did a study on Integrated Machine Learning and Bioinformatics Approaches for Prediction of Cancer-Driving Gene Mutations. In this study, several cancer-specific predictive models for prediction of driver mutations in cancer-linked genes were compared and were validated on canonical data sets of functionally validated mutations and applied to a raw cancer genomics data (Odeyemi, 2020). By analyzing pathogenicity prediction and conservation scores, showed that evolutionary conservation scores play a pivotal role in the classification of cancer drivers and were the most informative features in the driver mutation classification. Through extensive comparative analysis with structure-functional experiments and multicenter mutational calling data from Pan Cancer Atlas studies, demonstrated the robustness of the models and addressed the validity of computational predictions (Johnson, Ben-Zion, Meng and Vernon, 2020). The performance of the models were evaluated using the standard diagnostic metrics such as sensitivity, specificity, area under the curve and F-measure. To address the interpretability of cancer-specific classification models and obtain novel insights about molecular signatures of driver mutations, he complemented machine learning predictions with structure-functional analysis of cancer driver mutations in several key tumor suppressor genes and oncogenes. Through the experiments carried out in this study, it was revealed that evolutionary-based features have the strongest signal in the machine learning classification VII of driver mutations and provide orthogonal information to the ensemble-based scores that are prominent in the ranking of feature importance.

Zhu, Xie, Han and Guo, (2020) did a review on the Application of Deep Learning in Cancer Prognosis Prediction. The review, revealed that Deep learning as a generic model, requires less data engineering, and achieves more accurate prediction when working with

large amounts of data. The application of deep learning in cancer prognosis has been shown to be equivalent or better than current approaches, such as Cox-PH. With the burst of multi-omics data, including genomics data, transcript omics data and clinical information in cancer studies. The review proved that deep learning would potentially improve cancer prognosis.

Quang and Pham, (2020) studied machine learning approaches for breast cancer survivability prediction. The study entailed development of machine learning methods to identify meaningful biomarkers for breast cancer survivability prediction after a certain treatment. They include applying feature selection methods on gene-expression data to derived gene signatures, where the initial genes are collected concerning the mechanism of some drugs used breast cancer therapies. In addition, it has been increasingly supported that, sub-network biomarkers are more robust and accurate than gene biomarkers (Quang & Pham, 2020). They proposed two network-based approaches to identify sub-network biomarkers for breast cancer survivability prediction after a treatment. They integrate gene-expression data with protein-protein interactions during the optimal subnet searching process and use cancer-related genes and pathways to prioritize the extracted sub-networks. The sub-network search space is usually huge and many proteins interact with thousands of other proteins (Zhu, Xie, Han & Guo, 2020). Heuristics was applied to avoid generating and evaluating redundant sub-networks. Experimental results showed their approaches were effective. The results revealed that potential gene signatures and sub-network biomarkers are biologically meaningful and can yield significantly high accuracy in predicting breast cancer outcomes after treatment.

Integrated genomic analysis for prediction of survival for patients with liver cancer using the cancer Genome Atlas. To evaluate the prognostic power of different molecular data in liver cancer. Cox regression screen and least absolute shrinkage and selection operator were performed to select significant prognostic variables. Then the concordance index was calculated to evaluate the prognostic power. For the combination data, based on the clinical cox model, molecular features that better fit the model were combined to calculate the concordance index. Prognostic models were built based on the arithmetic summation of the significant variables. Kaplan-Meier survival curve and log-rank test were performed to compare the survival difference. Then a heat map was constructed and gene set enrichment analysis was performed for pathway analysis (Cammarota et al, 2020). The mRNA data were the most informative prognostic variables in all kinds of omics data in liver cancer, with the highest concordance index (C-index) of 0.61. For the copy number variation, methylation and

miRNA data, the combination of molecular data with clinical data could significantly boost the prediction accuracy of the molecular data alone ($P < 0.05$). On the other hand, the combination of clinical data with methylation, miRNA and mRNA data could significantly boost the prediction accuracy of the clinical data itself ($P < 0.05$). Based on the significant prognostic variables, different prognostic models were built. In addition, the heat map analysis, survival analysis, and gene set enrichment analysis validated the practicability of the prognostic models. In all kinds of omics data in liver cancer, the mRNA data might be the most informative prognostic variable. The combination of clinical data with molecular data might be the future direction for cancer prognosis and prediction (Cammarota et al, 2020).

Exploring cancer genomic data from the cancer genome atlas project catalogue of molecular aberrations that cause ovarian cancer is critical for developing and deploying therapies that will improve patients' lives. The Cancer Genome Atlas project has analyzed messenger RNA expression, microRNA expression, promoter methylation and DNA copy number in 489 high-grade serous ovarian adenocarcinomas and the DNA sequences of exons from coding genes in 316 of these tumors (Zhu, Xie, Han & Guo, 2020). Here we report that high-grade serous ovarian cancer is characterized by TP53 mutations in almost all tumors (96%); low prevalence but statistically recurrent somatic mutations in nine further genes including NF1, BRCA1, BRCA2, RB1 and CDK12; 113 significant focal DNA copy number aberrations; and promoter methylation events involving 168 genes. Analyses delineated four ovarian cancer transcriptional subtypes, three microRNA subtypes, four promoter methylation subtypes and a transcriptional signature associated with survival duration, and shed new light on the impact that tumors with BRCA1/2 (BRCA1 or BRCA2) and CCNE1 aberrations have on survival (Zhu, Xie, Han & Guo, 2020). Pathway analyses suggested that homologous recombination is defective in about half of the tumors analyzed, and that NOTCH and FOXM1 signaling are involved in serous ovarian cancer pathophysiology (Johnson, Ben-Zion, Meng and Vernon, 2020).

2.3.5 Existing cancer prediction models

Tsai *et al*., (2021) in their study which aimed to develop a Prediction Model for assessing individual metachronous Peritoneal Carcinomatosis (mPC) in Patients with Stage T4 Colon Cancer. A total of 2003 patients with pT4 colon cancer undergoing R0 resection were categorized into the training or testing set. Based on the training set, 2044 Cox prediction models were developed. Next, models with the maximal C-index and minimal

prediction error were selected. The final model was then validated based on the testing set using a time-dependent area under the curve and Brier score, and a scoring system was developed (Tsai *et al*., 2021). Based on the CART, patients were categorized into the low-risk or high-risk groups. The model had high predictive accuracy (prediction error ≤5%) and good discrimination ability (area under the curve >0.7). The study concluded that the prediction model quantifies individual risk and is feasible for selecting patients with pT4 colon cancer who are at high risk of developing mPC (Tsai *et al*., 2021). The model is limited since it is limited to colon cancer prediction.

Many state-of-the-art deep learning techniques have been applied to cancer prognosis prediction, indicating the great potential and the urgent need of utilizing multi-omics data from cancer patients to test new algorithm and improve model performance. There key challenges in applying deep learning approach in cancer prognosis prediction to achieve high performance.

The ANNs model is also subject to significant challenges. One of the challenges in using ANNs is mapping how the real-world input/output (an image, a physical characteristic, a list of gene names, a prognosis) can be mapped to a numeric vector. In ANNs the adjustment of neural connection strengths is usually done via an optimization technique called back-propagation. The ANNs model is also subject to significant challenges. One of the challenges in using ANNs is mapping how the real-world input/output (an image, a physical characteristic, a list of gene names, a prognosis) can be mapped to a numeric vector. In ANNs the adjustment of neural connection strengths is usually done via an optimization technique called back-propagation.

Javier *et al.,* (2021) in their study of developing and validating an individualized breast cancer risk prediction model for women attending breast cancer screening, used partly conditional Cox proportional hazards regression to estimate the adjusted hazard ratios (aHR) and individual risks for age, family history of breast cancer, previous benign breast disease, and previous mammographic features. They validated the model with the expected-to-observed ratio and the area under the receiver operating characteristic curve. All three risk factors were strongly associated with breast cancer risk, with the highest risk being found among women with family history of breast cancer (aHR: 1.67), a proliferative benign breast disease (aHR: 3.02) and previous calcifications (aHR: 2.52) (Javier *et al.,* 2021). The study concluded with a development of a risk prediction model to estimate the short- and long-term

risk of breast cancer in women eligible for mammography screening using information routinely reported at screening participation.

Lee *et al*., (2019) in their study of BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. They extend the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) risk model to incorporate the effects of polygenic risk scores (PRS) and other risk factors (RFs). The results indicated that the predicted lifetime risks for women in the UK population vary from 2.8% for the 1st percentile to 30.6% for the 99th percentile, with 14.7% of women predicted to have a lifetime risk of ≥17–<30% (moderate risk according to National Institute for Health and Care Excellence [NICE] guidelines) and 1.1% a lifetime risk of ≥30% (high risk) (Lee *et al*., 2019). The study concluded that the model enabled high levels of BC risk stratification in the general population and women with family history, and facilitate individualized, informed decision-making on prevention therapies and screening. A challenge in the implementation of the model is that its predictive accuracy is lower compared to other ML models. This is evident in chang *et al*., study which showed that the Predictive accuracy reached 90.17% using ML-adaptive boosting and 89.32% using ML-Markov chain Monte Carlo generalized linear mixed model versus 59.31% with BOADICEA for the Swiss clinic-based sample. The model is limited by the fact that does not account for risk factors associated with reproductive history and hormonal exposure and has limited utility in cases with small family history.

Hill *et al*., (2013), in their study of improving prostate cancer detection in veterans through the development of a clinical decision rule for prostate biopsy, they classified PCa stages differently and built their two models accordingly. In the first model, the difference in the discrimination was analysed and based on all PCa versus non-cancerous prostate conditions where the AUC for this model was 0.68 compared with 0.59 for PSA alone. In the second model, the discrimination analysis was based on PCa stages II, III, IV versus PCa stage I, prostatic interstitial neoplasm, BPH and prostatitis where stages I, II, III and IV are parallel to T1, T2, T3/T4 and metastatic PCa, respectively (Hill *et al*., 2013). The AUC for the second model was 0.72 compared with 0.63 for PSA alone. The study concluded that evaluating certain common biomarkers in conjunction with PSA may improve PC prediction prior to biopsy. Moreover, the biomarkers may be more helpful in detecting clinically relevant PC.

Babain *et al*., (2000) in their study which aimed to explore the potential role of a neural network-derived algorithm in enhancing the specificity of prostate cancer detection compared with the determination of prostate-specific antigen (PSA) and free PSA (fPSA) while maintaining a 90% detection rate. Performance parameters (including sensitivity, specificity, positive and negative predictive values, and biopsies saved) were calculated, and a comparative analysis was performed to evaluate the differences among the new algorithm, percent fPSA, PSA density, and PSA density-transition zone. Cancer was histologically confirmed in 24.5% (37 of 151) of the men under study. The median age of the men was 62 years (range 43 to 74) (Babain *et al*., 2000). At a sensitivity of 92%, the specificity for percent fPSA was 11%. The new algorithm (PCD-I) demonstrated an additional enhancement of specificity to 62% at 92% sensitivity. Clinically, the PCD-I would result in a savings of 49% (74 of 151) of all biopsies or 63.6% (71 of 114) of all unnecessary biopsies. The study concluded that a new generation algorithm, derived from a neural network (PCD-I) incorporating the parameters of age, creatinine kinase, PSA, prostatic acid phosphatase, and fPSA can significantly enhance the specificity and reduce the number of biopsies while maintaining a 92% sensitivity rate.

Salem *et al*., (2021) in their study which aimed to systematically review the applications of ES in urological research and their methodological models for effective multi-variate analysis, they applied the PRISMA methodology to formulate an effective method for data gathering and analysis. A total of 168 systems were finally included and systematically analysed demonstrating a recent increase in uptake of ES in academic urology in particular artificial neural networks with 31 systems. Most of the systems were applied in urological oncology (prostate cancer = 15, bladder cancer = 13) where diagnostic, prognostic and survival predictor markers were investigated. Due to the heterogeneity of models and their statistical tests, a meta-analysis was not feasible. The study concluded that ES utility offers an effective ML potential and their applications in research have demonstrated a valid model for multi-variate analysis (Salem *et al*., 2021). The complexity of their development can challenge their uptake in urological clinics whilst the limitation of the statistical tools in this domain has created a gap for further research studies. Integration of computer scientists in academic units has promoted the use of ES in clinical urological research. A systematic review of the applications of Expert Systems (ES) and machine learning (ML) in clinical urology. ES systems are limited by its difficulty in knowledge acquisition (Salem *et al*., 2021).

Mavaddat (2010) in their study of incorporating tumour pathology information into breast cancer risk prediction algorithms, extended BOADICEA by treating breast cancer subtypes as distinct disease end points. Age-specific expression of phenotypic markers in a series of tumours from 182 BRCA1 mutation carriers, 62 BRCA2 mutation carriers and 109 controls from the Breast Cancer Linkage Consortium, and over 300,000 tumours from the general population obtained from the Surveillance Epidemiology, and End Results database, were used to calculate age-specific and genotype-specific incidences of each disease end point. The probability that an individual carries a BRCA1 or BRCA2 mutation given their family history and tumour marker status of family members was computed in sample pedigrees (Mavaddat, 2010). The predicted BRCA1 carrier probabilities among ER-positive breast cancer cases were less than 1% at all ages. For women diagnosed with breast cancer below age 50 years, these probabilities rose to more than 5% in ER-negative breast cancer, 7% in TN disease and 24% in TN breast cancer expressing both CK5/6 and CK14 cytokeratins. Large differences in mutation probabilities were observed by combining ER status and other informative markers with family history. The study concluded that prediction of BRCA1/2 carrier status, and hence selection of women for mutation screening, may be substantially improved by combining tumour pathology with family history of cancer (Mavaddat, 2010). The model is limited by the fact that does not account for risk factors associated with reproductive history and hormonal exposure and has limited utility in cases with small family history.

2.4. Type of Cancer

The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci: 1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability. In the first case, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease. In the second case one is trying to predict the likelihood of redeveloping cancer after to the apparent resolution of the disease. In the third case one is trying to predict an outcome (life expectancy, survivability, progression, tumor-drug sensitivity) after the diagnosis of the disease. In the latter two situations the success of the prognostic prediction is obviously dependent, in part, on the success or quality of the diagnosis. However a disease

prognosis can only come after a medical diagnosis and a prognostic prediction must take into account more than just a simple diagnosis.

The type of cancer is a significant factor since a cancer prognosis typically involves multiple physicians from different specialties using different subsets of biomarkers and multiple clinical factors, including the age and general health of the patient, the location and type of cancer, as well as the grade and size of the tumor.

2.5 Gender

It is well known that men and women differ in terms of cancer susceptibility, survival and mortality, but exactly why this occurs at a molecular level has been poorly understood. Sex is a biological parameter that influences the development and progression of various diseases, including cancer. Sex and gender are often used interchangeably, but while sex refers to biological characteristics, gender can be defined as roles, behaviors, activities and attributes that a society considers suitable for male versus female from a cultural point of view. In fact, a gender perspective in health should also focus on people's circumstances in relation to their economic, social, cultural and working conditions. These are significative determinants that may impact on development, diagnosis and response to therapy.

Several gender related risk factors, such as smoking and other habits such as alcohol intake, sunlight exposure, environmental exposures, body weight, dietary patterns and physical activity behaviors have a different impact on men and women in this context. For most types of cancer, males show a higher risk of malignancy during their lifetime than females and have a worse prognosis. Cancer in females must evade more efficient immune surveillance mechanisms and undergo a more intense immune-editing process to become metastatic. This ability of tumors in females to evade immune surveillance makes metastatic tumors less immunogenic and enriched with more efficient immune escape mechanisms and may therefore exhibit resistance to immunotherapy.

2.6 Tumor Stage

Tumor size is a critical clinical factor with considerable prognostic and predictive value for T1 breast cancer, and it should be selectively incorporated into the current staging system to facilitate prediction of death and recurrence risk. Staging is a way to describe a

cancer. The cancer's stage tells you where a cancer is lojcated and its size, how far it has grown into nearby tissues, and if it has spread to nearby lymph nodes or other parts of the body. Before starting any cancer treatment, doctors may use physical exams, imaging scans, and other tests to determine a cancer's stage. Staging may not be completed until all the tests are finished.

Stratified analyses based on T stage found that the increase of T stage significantly and negatively repressed the effect of tumor size on death and recurrence risk. In addition, tumor size showed the greatest hazard ratio of cancer-specific death and relapse in T1 colon cancer. Even more importantly, the discriminatory ability of tumor size outperformed any other widely accepted prognostic clinical features in predicting cancer-specific survival

The proposed model can achieve a higher performance on cancer tumor classification using gene expression data. Both deep and machine learning methods and a combination of both supervised and unsupervised learning can assist in predicting or detecting cancer susceptibility in the early stages and therefore, aid in designing early treatment strategies, and in turn increase survival of the high-risk women.

2.7 Age

Age, defined by completed units of time,[1] is used in virtually all studies of cancer epidemiology and is one of the most studied risk factors for cancer. Cancer can be considered an age-related disease because the incidence of most cancers increases with age, rising more rapidly beginning in midlife. Age also can be considered a surrogate measure for the complex biological processes associated with aging. However, aging, the process of getting older, can be distinguished from age-associated diseases. Paradoxically, adults with the longest longevity are less likely to develop cancer. Thus, aging can be viewed as a natural process, not pathology, and old age does not necessarily lead to cancer. The risk of receiving a diagnosis of different types of cancer varies throughout a person’s life span. The cumulative risk for all cancers combined increases with age, up to age 70 years then decreases slightly.

Supervised and unsupervised learning models can be used to examine the influence of prenatal and early life events on cancer development in adulthood. A recent federal, interagency report on breast cancer research, for example, highlighted evidence that exposures that cause molecular and cellular changes in mammary tissue during puberty or

earlier can influence breast cancer development many years later. The finding that breast cancer incidence rates fell after the decline in the use of hormone replacement therapy at menopause suggests that critical periods for breast cancer development also exist later in life. In addition, opportunities may exist to intervene at midlife to alter or reverse disease processes that were initiated at earlier life stages.

2.8 Hereditary Factor

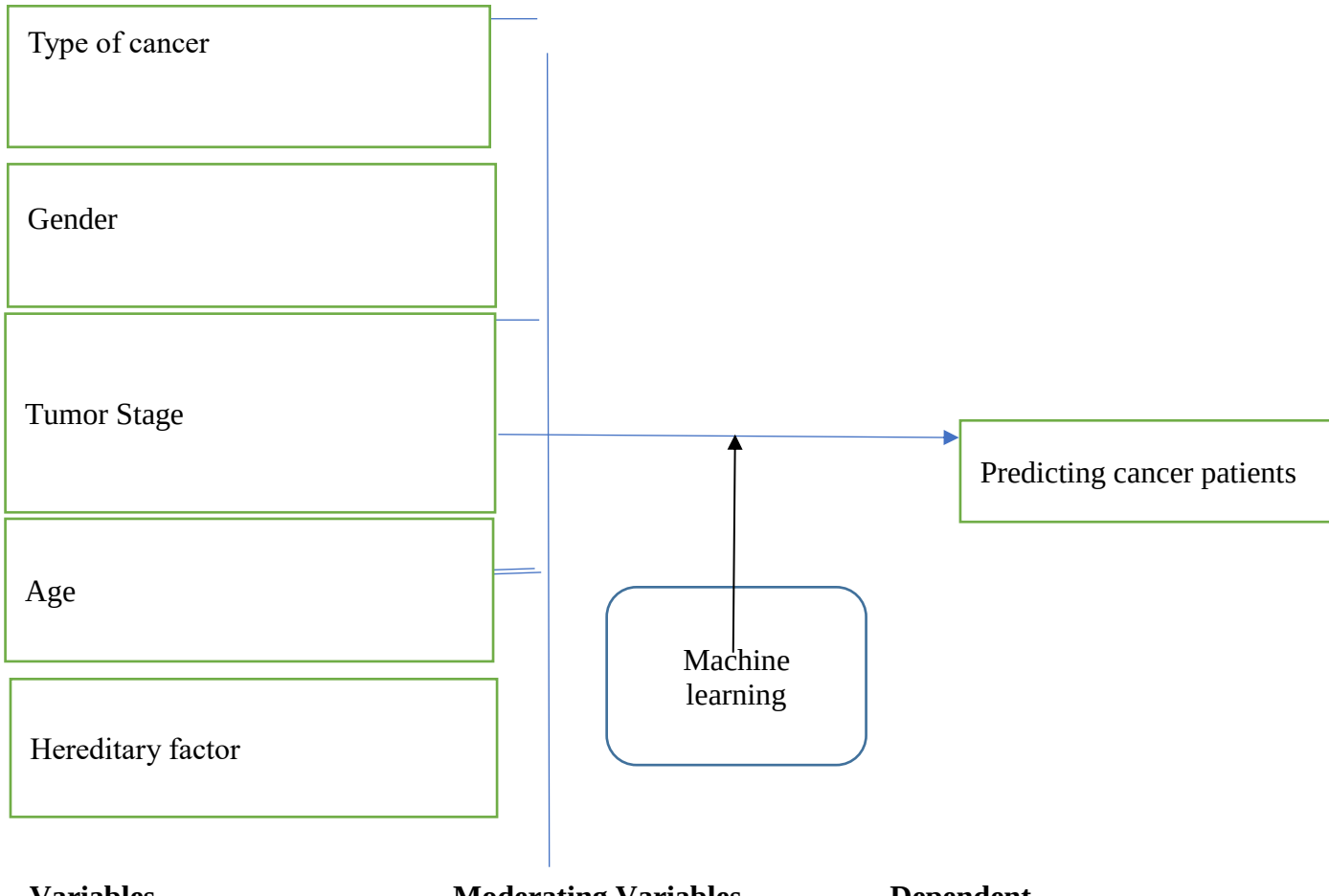
Cancer is a genetic disease that is, cancer is caused by certain changes to genes that control the way our cells function, especially how they grow and divide. Genes carry the instructions to make proteins, which do much of the work in our cells. Certain gene changes can cause cells to evade normal growth controls and become cancer. For example, some cancer-causing gene changes increase production of a protein that makes cells grow. Others result in the production of a misshapen, and therefore non-functional, form of a protein that normally repairs cellular damage.

Cancer-causing genetic changes can also be acquired during one's lifetime, as the result of errors that occur as cells divide or from exposure to carcinogenic substances that damage DNA, such as certain chemicals in tobacco smoke, and radiation, such as ultraviolet rays from the sun. Genetic changes that occur after conception are called somatic (or acquired) changes. In general, cancer cells have more genetic changes than normal cells. But each person's cancer has a unique combination of genetic alterations. Some of these changes may be the result of cancer, rather than the cause. As the cancer continues to grow, additional changes will occur. Even within the same tumor, cancer cells may have different genetic ch

2.9 Conceptual Framework

Figure 2.1

The Conceptual Framework



The conceptual framework highlights the role of machine learning and the use of gender, age, tumor stage, type of cancer, and hereditary factor in predicting cancer patients.

2.10 Operationalization of Variables

Operationalization is the process of defining variables into measurable factors. The entire process, allows different concepts to be measured both quantitatively and empirically. Exact definitions of variables is also set, improving on the quality and robustness of the design.

Variables	Sub-variables	Indicators	Values (data)
Type of cancer	Breast cancer Cervical cancer Prostate Esophageal Colorectal	Tissue type Location of tumor in the body	Test data
Gender	Male Female	Type of cancer	Gender statistics
Tumor Stage	First Second Third Fourth	Tumor (T) Node(N) Metastasis (M)	Cancer data
Age	Young Elderly	Skin surface Topography Eyes	Cancer data
Predicting Cancer Patients	Type of cancer Gender Tumor stage Age	Tumor stage Gender Age	Cancer Data

Table 2.1

Operational Variables

2.11 Summary

In the not-so-distant past, data generation was the bottleneck, now it is data mining, or extracting useful biological insights from large, complicated datasets. In the past decade, technological advances in data generation have advanced studies of complex biological phenomena. Machine learning has made remarkable contributions in bioinformatics, detecting cancer, creating new drugs, analyzing traffic patterns. However, it becomes difficult when data grow complex and huge, but the rise and advancement of machine learning algorithms have made it possible to solve many problems. This study will be used to develop a machine learning model for predicting cancer patients using integrated genomic data.

Artificial intelligence has empowered research in the healthcare sector. The availability of open-source healthcare datasets has motivated the researchers to develop applications which helps in early diagnosis and prognosis of diseases. Advancement in genome sequencing technology has empowered researchers to think beyond their imagination. Researchers are trying their hard to fight against various genetic diseases such as cancer. The next generation sequencing technologies have allowed researchers to screen changes at varying biological scales, such as genome-wide genetic variation, gene expression and small RNA abundance, epigenetic modifications, protein binding motifs, and chromosome conformation in a high-throughput and cost-efficient manner.

Although previous studies for profiling of circulating miRNAs that have been proven a promising class for cancer detection and identification. The methods suffer from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumors. Other researchers, have generated massive amounts and multiple types of big data to understand different biological phenomena with regards to cancer prognosis. However, poor progress has be made in the application of gene signatures in cancer prognosis.

CHAPTER THREE

METHODOLOGY

3.1. Introduction

This chapter sets out the procedure that will be followed in completing the study. In this section procedures and techniques that will be used in the collection, processing and analysis of data will be identified. Specifically, the following subsections are included; research design, target population, sampling and sampling procedure, research instrument, validity and reliability of the instrument, data collection procedures, data processing and analysis, research schedule and resources and budget.

3.2 Research design

The research was in the form of a cross sectional study based on its focus on analysing data from a representative sample. Machine Learning algorithms may be supervised, unsupervised or semi supervised (Bonaccorso, 2017). In this study a supervised learning algorithm was used. The data pre-processing was done first which helped to eradicate the noises from the data. The null or missing data will then be identified and they will be deleted. The pre-processed data was split into testing and training set. The training set is fed to the algorithm and they helped the algorithm to learn. The test set is used for testing the accuracy. Once a classification model was obtained using one or more ML techniques, classifier's performance was estimated. The performance analysis of each proposed model was 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 were used for assessing the overall performance of a classifier (Mohammed, Khan and Bashier, 2016). The classification algorithms that were used will be support vector machines (SVM) using the software the SVM program with the polynomial kernel of degree 2 in the LIBSVM package. 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.

3.3. Target population

The researcher utilized the Kenyatta National Hospital dataset that includes 44,000 cancer patients. This formed the target population used in the study. The population was narrowed down to 1,172 new cases between January of 2017 and June 2019. This was achieved through stratified sampling where the population was divided into strata based on the type of cancer. Ngechu (2004) defined a population as a well-defined or set of people, services, elements, events, group of things or households that are being investigated. In this study, the population of interest will be cancer patients in Nairobi. However, Mugenda and Mugenda, (2003) explain that the target population should have some observable characteristics, to which the researcher intends to generalize the results of the study.

3.4. Research Instrument

The research relied on data from the KNH cancer database. Information was abstracted from the data obtained which was used for assessing breast cancer patients. The information abstracted included patient's age, sex, and origin, type of cancer, and method of cancer diagnosis and year of diagnosis. The questionnaires were used since adequate information for the study was obtained. Information that would not have been given out had interviews been used will also be obtained.

The data collected sought to gather literature on factors to predict breast cancer patients using integrated genomic data. Additionally, it sought to understand how to establish regression model for predicting breast cancer patients using integrated genomic data. Additionally, the data was used to test and validate the regression model for predicting breast cancer patients using integrated genomic data.

3.5. Validity and Reliability of the instrument

Validity indicates the degree to which an instrument measures what it is supposed to measure; the accuracy, soundness and effectiveness with which an instrument measures what it is intended to measure (Kothari, 2004) or the degree to which results obtained from the

analysis of the data actually represent the phenomena under study (Mugenda & Mugenda, 2007). There are various types of validity; content, construct and criterion validity. Content validity indicates the extent to which items adequately measure or represent the content of the property or trait that the researcher wishes to measure. To ensure this kind of validity was achieved, the subject matter expert such as supervisors was ensured that they review the development of research instruments.

3.6. Data collection procedure

Data to be used in the study was obtained from uploaded files in the openML site. The information that was of consideration from the uploaded files included Patient's age, sex, origin and cancer type. The openML site was the most convenient data source for all the information targeted. Comparisons between ML versus BRCAT were based on performance assessment on five datasets: Simulated data and retrospective data from the Kenyan population.

3.7. Data processing and analysis

In reviewing the literature, assessing the details and testing and validating the model for predicting cancer patients using integrated genomic data machine learning model was applied. The data samples constituted the basic components. Every sample was described with several features and every feature consists of different types of values. 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 was also improved. In addition, in order to make the raw data more suitable for further analysis, pre-processing steps were applied that focus on the modification of the data. ML algorithms work better when the dimensionality is lower. Additionally, the reduction of dimensionality eliminated irrelevant features, reduce noise and 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 will allow for gathering the

described benefits of dimensionality reduction. This will be essential in calculating absolute lifetime risk of invasive breast cancer according to the regression model for specific race/ethnic groups and age intervals for each individual in the datasets. The analysis for each objective are summarized in table 3.1 below.

Table 3.1

Analysis of each Objective

Objectives	Analysis
To review the literature on factors to predict breast cancer patients using integrated genomic data.	Inferential data analysis was used in review of the literature based on its ability to test different ML theories.
To develop a regression model for predicting breast cancer patients using integrated genomic data.	Regression analysis was used to assess the interactions between genomic variants and BC risk factors.
To test and validate the regression model for predicting breast cancer patients using integrated genomic data.	Regression analysis was employed as a predictive modeling technique that assesses the affiliation between the variables.

3.8 Model Development Process

The research employed regression analysis as the predictive modelling technique that assesses the affiliation among two or more variables. Regression analysis was focused on the affiliation between the dependent and independent variables. The research was in the form of linear regression that finds the linear relationship between the dependent variable and independent variables using a best-fit straight line. Generally, a linear model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term (also called the intercept term). In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and the nature of the regression line is linear.

In this case, it sought to establish the ability of type of cancer, gender, tumor stage, and familial history through ML can facilitate the prediction of B.C. The variables fit the regression analysis since is used with naturally-occurring variables, rather than variables that have been manipulated through experimentation (Mitchell, 2019). In this case, the researcher relied on decision trees that are a form of supervision learning algorithm that splits the sample based on particular questions regarding the sample. The use of decision trees was

considered vital for the classification of problems (Mitchell, 2019). Additionally, the researcher used it since it is easy to understand and effective. The technique was vital in defining the most significant variables and highlighting the relation between two or more variables.

Additive and multiplicative models are two classical ML approaches for modelling the effect of multiple factors on disease (Zitnik, et al., 2019). Both approaches are based on regression methods; in additive models, the risk of disease has an additive form that generally uses linear regression, while multiplicative models use logistic regression to report the relative risk or odds ratio (OR). Using a multiplicative approach, the breast analysis of disease incidence and carrier estimation algorithm (BOADICEA) were developed to identify high-risk women based on known genetic and non-genetic risk factors, including information on BC pathology, demographi factors, and variants of high-risk genes.

Although the BOADICEA model has been validated with large-cohort data, its discriminatory power in identifying high-risk women is limited. The model assumes that risk factors are independent of each other and interact in a linear way with BC development. Feld et al. (2007) also evaluated the predictive performance of combinations of 4 demographic risk factors, 10 published BC risk-associated SNPs, and 4 mammography features to predict BC risk in a case-control study with four logistic regression models. They showed that a combination of data improves BC risk prediction over methods that use only a subset of features. However, one should note that these studies are often based on a limited number of predictor variables and conventional regression models, which might make the estimates imprecise when working with potential multicollinearity in high-dimensional medical data, such as in genetic variants, To address this knowledge gap.

In this study, the researcher adopted the ML approach previously published in Behravan et al. (2018), which is built on an extreme gradient tree boosting (XGBoost) model followed by adaptive iterative feature selection, to capture optimal networks of interacting features (genetic variants and demographic risk factors for BC) in a BC risk prediction task.

The Extreme Gradient Boosting (XGBoost) is a new tree-based algorithm that has been increasing in popularity for data classification recently, that has been proved to be a highly effective method for data classification. The XGBoost is a highlyscalable end-to-end tree boosting system used in machine learning for classification and regression tasks (Chen &

Guestrin, 2016). The researcher replaced the Fully Connected Layer (FCL) from the DenseNet201 with the XGBoost classifier. This is because the original FCL classifies on the ImageNet dataset which consists of non-medical related images. The authors that proposed this method, Chen and Guestrin have explained their concept of approach in detail. This method is new, the researcher summarized the calculations and definitions.

First, a tree ensemble method of classification and regression trees (CARTs) with a set of $Ki/E|i \in 1 \ldots K$ nodes. The final prediction output of class label $\hat{y}i$ is calculated based on the total prediction scores at a leaf node fk for each tree kth. As expressed below.

$$\hat{\boldsymbol{y}}\boldsymbol{i} = \boldsymbol{\varphi}(\boldsymbol{x}\boldsymbol{i}) = {}^{K}\sum\nolimits^{k=1} \boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{i}}),\ \boldsymbol{f}_{\boldsymbol{k}} \in \boldsymbol{F},$$

where $\boldsymbol{x}_{\boldsymbol{i}}$ is the training set and F represents the set of all K scores for all CARTs. Then, a regularization step is applied to improve the results, as shown below.

$$\mathbf{L}(\boldsymbol{\varphi}) = \sum \boldsymbol{i}\mathbf{l}(\hat{\boldsymbol{y}}\boldsymbol{i},\boldsymbol{y}\boldsymbol{i}) + \sum \boldsymbol{k}\ \boldsymbol{\Omega}(\boldsymbol{f}\boldsymbol{k}),$$

where lrepresents the differentiable loss function, define by calculating the error difference between target yi and predicted class labels $\hat{y}i$. The second part performs penalization Ω on the model complexity to avoid over-fitting problems. The function for the penalty Ω is calculated below

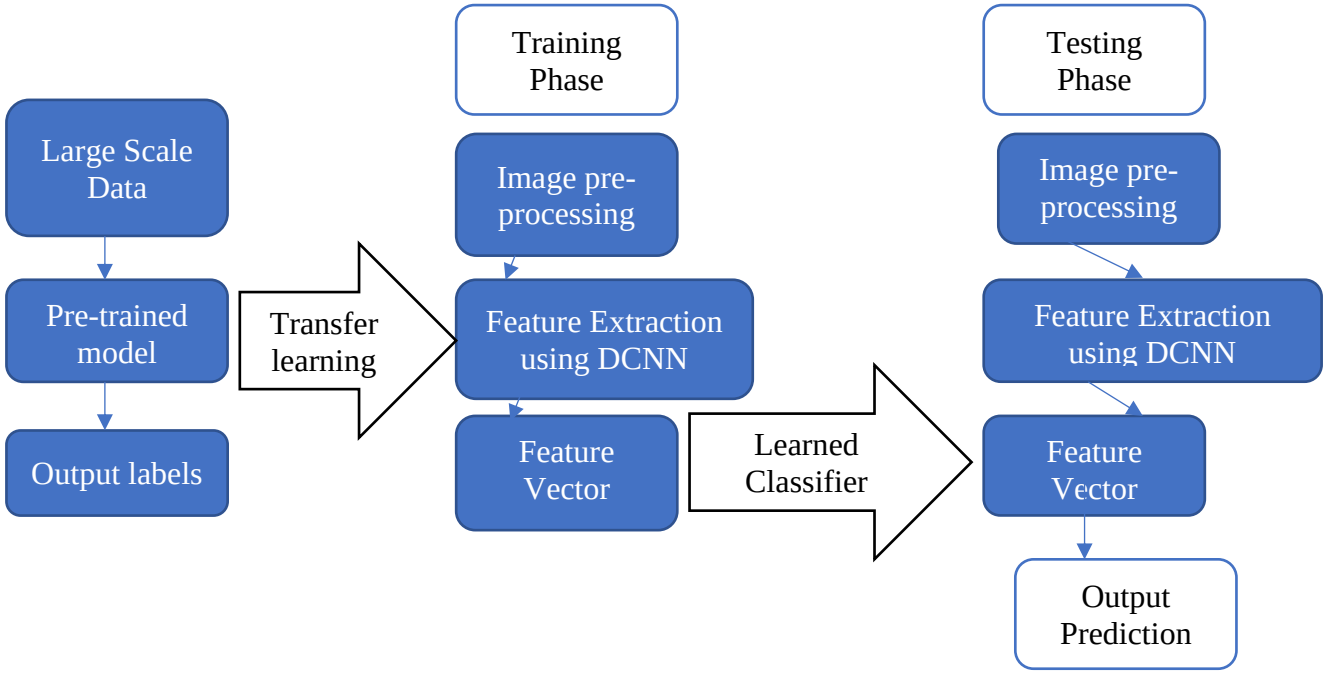
$$\boldsymbol{\Omega}(\boldsymbol{f}) = \boldsymbol{\gamma}\boldsymbol{T} + \mathbf{1}/\mathbf{2}\boldsymbol{\lambda}\ {}^{T}\sum\nolimits_{j=1} \boldsymbol{w}\mathbf{2}_{\boldsymbol{j}}$$

where γ and λ are configurable parameters to control the degree of regularization. T represents the leaves in the tree and w stores the value of weights for each leaf.

Then, Gradient Boosting (GB) is applied to effectively solve the classification problem along with the loss function and extended by a second Taylor expansion. The constant term will be removed to obtain a simplified objective at step t.

Figure 3.1

Overview of Proposed Method



An iterative process was used to identify the best combination of factors of BC. The data does not overlap between training, validation and test subset.

Input: K_1, K_2, where K_1 is the number of iterations and K_2 inner folds.
Input: S, data containing SNP features and target Y.
Input: G, data containing demographic features and target Y.
Required: Function **I**, adaptive iterative search algorithm to find the best interacting features (from[16]).
1. for $i = 1$ **to** K_1 iterations **do**
2. Randomly **split** S into S_i^{train}, S_i^{test} and G into G_i^{train}, G_i^{test} for i'th split with 80:20 split ratio.
3. **for** $j = 1$ to K_2 splits **do**
4. Shuffle and split S_i^{train} into S_j^{train}, $S_j^{validation}$ and G_i^{train} into G_j^{train}, $G_j^{validation}$ for j'th split.
5. Run **I** on S_j^{train} and $S_j^{validation}$ to identify the most important subset T of interacting BC risk-predictive SNPs.
6. Concatenate subset T of S_j^{train} with G_j^{train} to form C_j^{train}.
7. Concatenate subset T of $S_j^{validation}$ with $G_j^{validation}$ to form $C_j^{validation}$.
8. Run **I** on C_j^{train} and $C_j^{validation}$ to identify the most important subset T^* of interacting genetic and demographic risk factors of BC.
9. Fit classifier M on C_j^{train} using subset T^* features.
10. Form C_i^{test} by concatenating S_i^{test} and G_i^{test}.
11. Compute model performance for M on C_i^{test} using subset T^* features.
12. Calculate average performance across test sets.

CHAPTER FOUR

RESEARCH FINDINGS AND DISCUSSION

4.1 Introduction

The chapter highlights the findings of the research. Its aim was to develop a regression model for predicting Breast Cancer patients using Integrated Genomic Data.

4.2 Demographic Information

The type of cancer, gender, tumor stage, age, and hereditary factors were evaluated for each patient. Additional factors that were assessed include pathological diagnosis and treatment methods .Also abstracted from the medical records were the age at onset of menarche for female patients, family history of breast disease and the initial method of establishing the diagnosis. The data were analysed to calculate frequencies, means and standard deviations.

A total of 1172 new cases were seen at the KNH breast clinic over the two-and-a-half-year period, an average of 469 patients per year or 11 new patients per clinic visit. The mean age of these patients was 34.71 years with a range of 1 to 96 years. Table 4.1 gives some descriptive demographic characteristics.

Table 4.1:

Demographic Characteristics

Variable	Number	Mean	Std. Dev.	Min.	Max.
Age	1172	34.71	15.829	1	96
Duration of Symptoms	920	6.76	13.616	0.1	120

Type of Cancer and Gender

During the report period, breast cancer was the most frequent cancer among females, closely followed by cancer of the cervix uteri. On the other hand, cancers of the head & neck followed by oesophagus and prostate, lead in frequency among males. The pattern of most common cancers varied slightly when one compares single years in the report period.

However, it is worth noting that cancers of the breast and cervix uteri comprise a large proportion (43.3%) of all reported cases.

Among men, prostate, oesophageal and colorectal are the leading cancers in incidence, and in women, breast, cervical and oesophageal cancers are the most common. The leading cause of cancer death in Kenya is oesophageal followed by cervical and breast cancer. It is further estimated that there are 3200 new cancer cases among children below 18 years with the top five commonest cancers being Leukaemia, Non-Hodgkin's lymphoma, kidney cancer, brain cancers and cancer of the naso-pharynx.

Age

This cross-sectional ML based analysis was conducted based on the BC dataset Kenyatta National Hospital. Since the data was obtained from the public domain, ethical clearance consent was not required for conducting this analysis. It included 9 independent variables and 1 dependent variable. The independent variables were: age, patient's age (in years) at the time of diagnosis, reported as 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79, menopause (menopausal status of the patient at the time of diagnosis, reported as pre-menopause, lt40 and ge40 (further details were not provided for lt40 and ge40), tumor size (the size of the tumor (in mm), reported as 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50- 54, and 55-59), invasive nodes (the number of lymph nodes showing BC at the time of histological examination, reported as 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, and 36-39), node-caps (the penetration (yes or no) of the tumor in the lymph node capsule), degree of malignancy.

Table 4.2:

Age

Variable		Recurrence		P value
		No (n=196).	Yes (n=81)	
Age	20-49	84 (42.9%)	42 (51.9%)	0.002169
	50-79	112 (57.1%)	39 (48.1%)	

Tumor Stage

The histological grade of the tumor, where grade 1: looks most like normal breast cells and is usually slow-growing; grade 2: looks less like normal cells and is growing faster and grade 3: looks different to normal breast cells and is usually fast-growing), breast (the breast (left or right) affected with BC), breast quadrant (the specific location of the breast affected with BC, reported as left-upper, left-lower, right-upper, right-lower and central); irradiation (the radiation therapy history of the patient (yes or no). The dependent variable was class (the recurrence status (yes or no) of the patient.

Table 4.3:

Tumor Stage

Variable		Recurrence		P value
		No (n=196).	Yes (n=81)	
Tumor Size	0-9	11 (5.6%)	1 (1.2%)	0.00218
	10-19	50 (25.5%)	7 (8.6%)	
	20-29	67 (34.2%)	32 (39.5%)	
	30 and above	68 (34.7%)	41 (50.6%)	

Hereditary Factor

Distribution of the BC risk factors related to familial history in the KNH dataset. The P-values denote the differences in the Group 1 features between the BC cases and controls using the chi-squared test for categorical variables. The difference is statistically significant when the p-value < 0.05 (highlighted p-values). P-values were not adjusted for multiple testing.

Table 4.4

Distribution of BC Risk Factors in Familial History

Features Names	Cases		Controls	All Subjects	Description	L;
Cancer in Family	0	196- 44%	136 (54%)	332 (48%)	Is there any cancer in family members: 0: No; 1: Yes	0.01
	1	249- 56%	114 (46%)	363 (52%)		
Cancer type 1	0	394- 89%	235 (94%)	629 (90%)	Type of cancer in the 1st family member with cancer: 0: Other; 1: Breast	0.02
	1	51- 11%	15 (6%)	66 (10%)		
Cancer type 2	0	437- 98%	249 (99%)	686 (99%)	Type of cancer in the 2nd family member with cancer: 0: Other; 1: Breast	0.2
	1	8- 2%	1 (1%)	9 (1%)		
First-degree relative 1	0	398- 89%	238 (95%)	636 (91%)	whether the 1st family member with breast cancer is a first-degree relative: 0: No; : Yes	0.01
	1	47- 11%	12 (5%)	59 (9%)		

‘

Lateral 1	0	394- 88%	235 (94%)	629 (90%)	Whether the 1st family member has unilateral or bilateral BC: 0: No tumour; 1: Unilateral; 2: Bilateral	
	1	48- 11%	14 (5%)	62 (9%)		
	2	3- 1%	1 (1%)	4 (1%)		
Lateral 2	0	437- 97%	249 (99%)	686 (97%)	Whether the 2nd family member has unilateral or bilateral BC: 0: No tumour; 1: Unilateral; 2: Bilateral	
	1	7- 2%	1 (1%)	8 (2%)		
	2	1%	0 (0%)	1 (1%)		

First-degree relative 2	0	438- 98%	249 (99%)	687 (99%)	whether the 2nd family member with breast cancer is a first-degree relative: 0: No; 1: Yes	0.3
	1	7- 2%	1 (1%)	8 (1%)		
No. of BCs	0	393- 88%	235 (94%)	628 (90%)	Number of family members with BC.	0.04
	1	45- 10%	14 (5%)	59 (9%)		
	2	7- 2%	1 (1%)	8 (1%)		
BC risk score	0	393- 89%	238 (95%)	634 (91%)	Number of first-degree family members with BC.	0.02
	1	44- 10%	11 (4%)	55 (8%)		
	2	5- 1%	1 (1%)	6 (1%)		

4.3 Research Findings

4.3.1 Factors to Predict Breast Cancer Patients using Genomic Data

The breast clinic's records and case files of all patients managed for breast diseases between 2017 to 2020 were retrospectively reviewed. The KNH breast clinic was set up in 2000 for the purpose of optimizing the care of breast diseases. It is held once a week and run by consultant surgeons and surgical residents in training. All patients seen at the clinic during the study period with a clinically and/or cytologically, histologically, radiologically diagnosed breast ailment were included in the study. Patients who presented with previously treated or recurrent lesions were excluded.

Malignant breast lesions accounted for 22% of all breast diseases. The most common malignancy was ductal carcinoma constituting 91.7% of all cancerous breast diseases followed by lobular carcinoma at 2.8% (Table 4.2).

Table 4.5:

Distribution of Malignant Lesions

Diagnosis	Frequency	%
Ductal carcinoma	231	91.7
Lobular carcinoma	7	2.78
Malignant phyllodes	6	2.38
Poorly differentiated carcinoma	3	1.19
Primary soft tissue sarcoma	2	0.79
Squamous cell carcinoma	2	0.79
Ductal carcinoma with pagets	1	0.4
Total	252	100

The majority of patients (78%), had benign conditions, fibroadenoma being the single most common diagnosis made. It accounted for up to 40.2% of benign conditions (Table 4.3) and 33.2% of all breast ailments put together. Five disease conditions: fibroadenoma, breast abscess, fibrocystic disease, mastalgia and chronic mastitis accounted for 83.4% of all benign lesions and formed the main work load at the clinic. Out of the 22 patients with chronic mastitis, eight had confirmed tuberculosis of the breast.

Table 4.6

Distribution of Benign Breast Cancer

Diagnosis	Frequency	%
Fibroadenoma	370	40.2
Abscess	173	18.8
Fibrocystic disease	119	12.9
Mastaglgia	83	9.02
Mastitis chronic	22	2.39
Duct ectasia	12	1.3
Galactocoele	12	1.3
Gynaecomastia	12	1.3
Tubular adenoma	10	1.1
Lipoma	10	1.1
Others	86	9.4
Inconclusive	11	1.2
Total	252	100

Males constituted less than 1% in this series and gynaecomastia was the most common lesion seen in this group (Table 4.4).

Table 4.7

Distribution of Male Breast Pathology

Diagnosis	Frequency	%
Gynaecomastia	7	36.8
Fibrocystic disease	2	10.5
Fibradenoma	3	15.8
Chronic abscess	1	5.3
Malignant phyllodes	1	5.3
Ductal Carcinoma	4	21.1
Inconclusive	1	5.3
Total	19	100

Fine needle aspiration cytology (FNAC) was used in the initial definitive diagnosis in 46.1% of all cases. Excisional and incisional biopsies were utilised for only 8.6 and 3.0% of cases respectively. Patients diagnosed purely on clinical assessment comprised 40.0% (Table 4.5).

Table 4.8

Mode of Initial Definitive Diagnosis

Diagnosis	Frequency	%
FNAC	414	46.1
Critical Examination	359	40
Excisional Biopsy	78	8.6
Incisional biopsy	27	3.0
Others	21	2.3
Total	899	100

Over 80% of patients seen at the KNH breast clinic were operated upon. The remainder population were reassured, got radiotherapy, non-cancer oral treatment, and chemotherapy.

Table 4.9

Primary Treatment Offered Initially

Treatment Modality Offered	Frequency	%
Surgery	903	81.1
Reassured	79	7.1
Radiotherapy	55	4.9
Non-cancer oral Medication	46	4.1
Chemotherapy	27	2.4
Others	4	0.4
Total	1114	100

4.3.2 Results for Regression Model for Predicting Breast Cancer Patients using Genomic Data

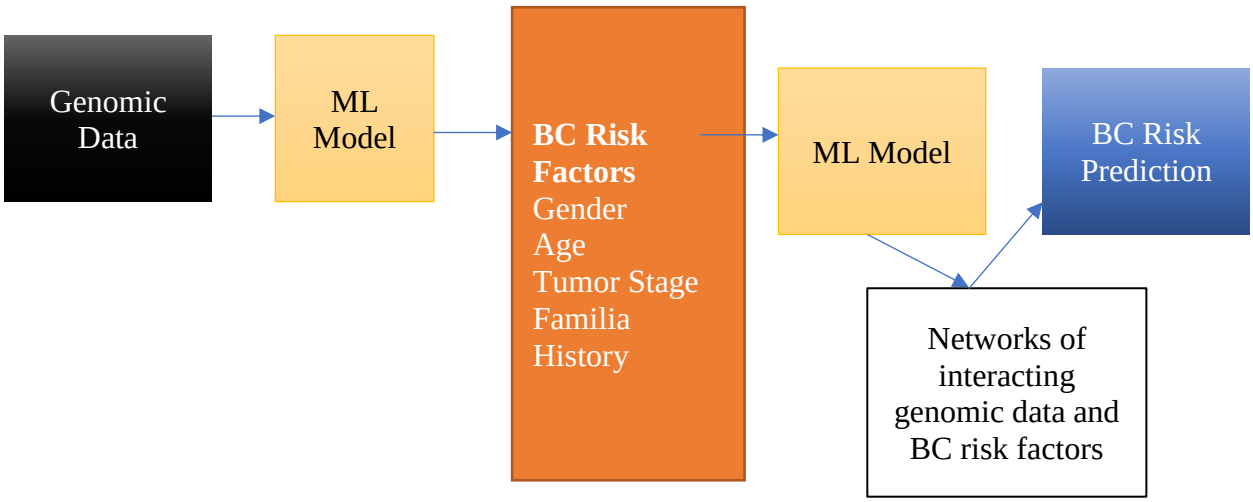
The statistical analysis of the extracted data was performed using R Project for Statistical Computing. For univariate analysis, the effect of each variable was tested for statistical significance using the chi-square test. All comparisons were two-tailed and $p<0.05$ were considered to be statistically significant. A multivariable analysis (predictive model development) was carried out using a logistic regression model where all the variables with $p<0.2$ in the univariate analysis were analyzed using an enter method. Before applying this multivariate model, the number of patients in the recurrence and non-recurrence group was

balanced using Synthetic Minority Over-sampling Technique (SMOTE), one of the most popular algorithms for balancing the dataset. The balanced dataset was split into training and test data in 70:30 ratios. The training dataset included both independent variables and a dependent variable (class) and was used to train the logistic regression model. While, the test dataset was used to assess how well the model was trained.

Figure 4.2 below illustrates the outline of the BC risk prediction system developed in this work. The ML model is trained to find the best groups of interacting genetic and demographic risk factors that contribute to BC risk prediction. The researcher proposes that a unified BC risk prediction system that takes advantage of the interactions among both the risk factors within a family of variables (e.g., genetic variants) and the risk factors in different families of variables (e.g., genetic variants and demographic features) is highly desirable in the BC risk evaluation task. Note that this study serves as an example showing how ML can combine different components of cancer risk for risk evaluation, and the proposed approach can be extended to other multifactorial diseases.

Figure 4.2

Proposed BC Risk Prediction Model Architecture.



The training phase of the model is where networks of interacting genetic and demographic risk factors for BC are identified. These networks of features are then used to predict whether an unlabelled individual is a cancer case or a healthy control in the testing phase. The model proves that a combination of interacting genomic data with BC risk factors related to both familial history, age, gender, and tumor stage increase BC risk prediction

accuracy. The researcher demonstrates the approach on the KNH dataset, which contains both genotyped data and demographic risk factors in 445 BC cases and 250 controls. The proposed system is compared with analyses based on only demographic risk factors for BC or on genetic variants. The researcher evaluated the approach against a model that combines 82 known BC-risk-associated SNPs and known demographic risk factors for BC.

4.4.3 Results for Test and Validation of the Model

The researcher demonstrates the approach on the KNH dataset, which contains both genotyped data and demographic risk factors in 445 BC cases and 250 controls. The proposed system is compared with analyses based on only demographic risk factors for BC or on genetic variants. The researcher evaluated the approach against a model that combines 82 known BC-risk-associated SNPs and known demographic risk factors for BC.

The performance of the model was assessed using standard statistical measures such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and AUC - Receiver Operating Characteristics (ROC) curve, where an AUC of 1.0 indicates perfect predictive ability, whereas 0.5 represents no predictive discrimination. The variables found to be statistically significant ($p<0.05$) in the multivariate logistic regression model were then used for a nomogram development, which provided the probability of BC recurrence. The total scores obtained from the nomogram were used to identify a cut-off score to categorize the BC patients into a high or low risk of recurrence. The performance of this cut-off score was assessed using accuracy, sensitivity, specificity, and AUC-ROC. Values ranging from 0.7 to 0.8 represent reasonable discrimination, and values exceeding 0.8 represent good discrimination.

The researcher also evaluated the model from a clinical point of view with pathologic T stage, pathologic N stage, subtypes according to ER/PR and HER2 status, EGFR status, and CK5/6 status by comparing the predicted recurrence proportion with the actual recurrence proportion. The Mean Absolute Error (MAE) and Weight Mean Absolute Error (wMAE) of each group showed great results of as little as 3.5%.

The model errors for pathologic T stage and N stage features were similar to each other but differed from those for the other pathologic features. The subtypes had similar error

values of around 2.5%. The discrimination of the wMAE at each prediction time (2, 5, and 7 years) showed only small differences.

4.6 Discussion of Results

The research established that demographic risk factors were also found to be individually important in our BC risk prediction model, although their importance scores were not equal. Similar to the genetic variants, combinations of demographic risk factors yielded a higher risk prediction accuracy than the individual demographic risk factors.

The same BC surveillance guideline since early 2010 were utilized in the research. Those guidelines recommend taking a careful history and performing a physical examination every 6–12 months, including regular mammography 6 months after the completion of definitive radiation therapy. In addition, the use of complete blood counts, chemistry panels, and tumor markers (CEA, CA-15-3) is not recommended for routine follow-up in an otherwise asymptomatic patient with no specific findings on clinical examination according to those guidelines. Understanding of the nature and biology of BC has improved, and it is now known that the timing and pattern of BC recurrence differ for patients with different BC subtypes. Moreover, the current concept of oligometastasis in BC, defined as low-volume metastatic disease with a few, small metastatic lesions, considered BC patients with oligometastatic to be a distinct subgroup with a more favorable long-term prognosis than patients with metastatic BC. This suggested that an early diagnosis of BC recurrence, rather than waiting for patients to show symptoms, might thus confer a survival benefit. Therefore, improved screening programs that incorporate the biology of individual BC patients and a method to precisely predict the risk of recurrence for individual patients are urgently needed.

This study reviewed data of 1172 patients seen at the KNH breast clinic over a two-and-a-half-year period. The majority, 98.9%, were females which is consistent with the rarity of male breast disease (prevalence ranges from 0 to 5.8%) in most series (1-3). Even then the majority of male breast afflictions are known to be benign, (2,4) as indicated by this study where 73.6% (14 of 19) were benign lesions with gynaecomastia being the commonest diagnosis made. In a retrospective study in Saudi Arabia over a 15-year period (n = 63), Chiedozi et al found that 87% of all male breast lesions were benign (1). The same study revealed that gynaecomastia was the most frequently diagnosed condition in males, 54%, and that only 3% of all breast cancers occurred in the male population.

Overall, fibroadenoma was the single most common diagnosis made followed by ductal carcinoma of the breast. This compares well with other studies where fibroadenoma shows prevalence rates between 34.7 and 67% of all breast lesions and a peak mean age incidence of 16-25 years (1-4,6, 10-13). Five conditions namely fibroadenoma, ductal carcinoma, breast abscesses, fibrocystic disease and breast pain (mastalgia) accounted for over 85% of all breast ailments seen. These five conditions share similar prominence in other studies, only differing in their order of ranking. Two studies, one in Jordan (n = 1000) and the other in Pakistan (n = 3879) found breast cancer to top the list of breast pathologies followed by fibrocystic disease and fibroadenoma in respective studies.

Fine needle aspiration cytology (FNAC) was the most commonly used initial pathological diagnostic investigation for breast lumps. This is tandem with worldwide trends, and the declining role of incisional and excisional biopsy in the initial diagnosis of solid and cystic masses of the breast. FNAC is easy to perform with minimal expertise even in the clinic setting, does not require anaesthesia and is less invasive compared to other methods. However, it requires well trained cytologists in order to reduce false positive or negative results. The correlation between cytology and histology in breast lesions has been studied at KNH and found to be good. Forty percent of lesions were diagnosed on the basis of clinical assessment alone. Presumably, breast abscesses, mastalgia and fibroadenomas in a young population, which comprised a large proportion of the current data set, informed the decisions.

Of the patients presenting with lumps, two thirds had masses greater than 5cm. This is consistent with the long duration of symptoms in the current study and the the earlier study that denoted that up to 70% of patients diagnosed with breast cancer have advanced disease at KNH. As an initial means of treatment, over 80% of patients seen were operated upon with only 7.1% getting reassurance only. This may seem an excessive initial treatment both for the majority fibroadenomas and the breast malignancies. These two conditions constituted 52.9% of all breast ailments. Most of the work load in any breast clinic is largely considered to be that of reassuring the “worried well”, implying that most pathologies encountered are benign, non–life threatening conditions that should be treated conservatively. Fiboadenomas are benign fibroepithelial tumors that grow slowly, are rarely symptomatic, tend to be bilateral and multiple, do not have. predilection to malignancy and up to 40% may regress within two years. It is unclear whether the decisions to excise them in the current study were influenced

by their large sizes (giant fibroadenomas), symptoms, sudden change in biology or patient request. For the advanced malignant lesions, the place of neoadjuvant therapy is now advocated. It is possible that this form of initial therapy was neither accessible nor available for the cohort. Further studies to evaluate our breast surgery practice are recommended.

In this study, demographic risk factors were also found to be individually important in our BC risk prediction model, although their importance scores were not equal. Similar to the genetic variants, combinations of demographic risk factors yielded a higher risk prediction accuracy than the individual demographic risk factors. Genomic and non-genomic entered into many predictive models as predictors, given that these factors represent breast cancer prediction. The tests used in the model were routinely performed at every follow-up visit. Thus, the machine learning model for BC prognosis was made with maximal use of the laboratory test results from current surveillance practices without requiring other laboratory work, such as intrinsic subtyping. Therefore, the BC prognosis model could fit into routine clinical practice better than previous machine learning models. Moreover, we can adapt this prognosis model into the EMRs using a website and thereby acquire information about BC recurrence in real time. This model could thus present the recurrence risk at each follow-up point using all available laboratory and imaging test results.

The proposed system was compared with analyses based on only demographic risk factors for BC or on genetic variants. The researcher evaluated the approach against a model that combines 82 known BC-risk-associated SNPs and known demographic risk factors for BC. The results show that merging genetic and non-genetic risk factors could enable the development of risk adapted screening programs, which can, in turn, categorize individuals based on their risk of developing cancer and then send those with a high risk of developing cancer for more precise screening e.g., by performing mammography, MRI and/or tumour segmentation. This could potentially improve the performance of BC screening and lead to an efficient allocation of clinical resources. This is in-line with literature by Gagnon, et al., (2016).

The results are aligned with research by Zhu, Xie, Han and Guo, (2020) that established that deep learning as a generic model are capable of improving cancer prognosis. The findings further support the literature by Tsai *et al*., (2021) that established that many state-of-the-art deep learning techniques have been applied to cancer prognosis prediction,

indicating the great potential and the urgent need of utilizing multi-omics data from cancer patients to test new algorithm and improve model performance.

4.5 Summary

The study findings established that demographic risk factors were also found to be individually important in our BC risk prediction model, although their importance scores were not equal. Similar to the genetic variants, combinations of demographic risk factors yielded a higher risk prediction accuracy than the individual demographic risk factors. The research findings further established that factors influencing breast cancer prognosis, screening appropriate predictors as independent variables are an important step in model construction. Age, disease stage, grade, tumor size, race, marital status, number of nodes, histology, number of positive nodes and primary site code have been entered into many predictive models as predictors, given that these factors represent key risk factors for onset and survival in breast cancer.

This chapter proposed an ML approach to efficiently combine genomic variants with BC risk factors and to search for optimal interactions among them. The proposed approach considerably increased the BC risk prediction accuracy compared to systems based solely on genetic variants or demographic risk factors for BC. To summarize, the main contributions of the present study are as follows:

i) identifying the networks of interacting genetic and demographic risk factors for BC that contribute most to predicting the BC risk,
ii) proposing an efficient ML framework to combine different risk factors for a multifactorial disease such as BC in a high-dimensional and partly small-sample-size problem,
iii) capturing non-linear interactions among the risk factors and modelling BC risk in a non-additive form.

In future, the results will help to create more effective ways to identify people at risk for BC, to whom screening methods should be directed. The proposed model is also adaptable to all other multifactorial disease entities.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATION

5.1 Introduction

This chapter presents an assessment of the data gathered in the survey and the findings obtained. The chapter presents the conclusions and recommendations from the research.

The proposed system is compared with analyses based on only demographic risk factors for BC or on genetic variants. The researcher evaluated the approach against a model that combines 82 known BC-risk-associated SNPs and known demographic risk factors for BC. The results show that merging genetic and non-genetic risk factors could enable the development of risk adapted screening programs, which can, in turn, categorize individuals based on their risk of developing cancer and then send those with a high risk of developing cancer for more precise screening.

5.2 Conclusion

The main objective of the study was to develop a regression model for predicting breast cancer patients using integrated genomic data. It was facilitated by the objectives that sought to review the literature on factors to predict breast cancer patients using integrated genomic data, develop a regression model for predicting breast cancer patients using integrated genomic data and test and validate the regression model for predicting breast cancer patients using integrated genomic data. Data was obtained online through openML site. Information was abstracted from the the data obtained which was used for assessing breast cancer patients. According to openML site, there are 44000 cancer patients, formed the target population. Analysis was conducted by reviewing the literature, assessing the details and testing and validating the model for predicting cancer patients using integrated genomic data machine learning model was applied. Inferential data analysis was used in reviewing the literature. In this case, the data was summarized into points in a constructive manner. Regression analysis was used in the identification of supervised learning models and their influence on the topic. Additionally, regression analysis was employed as a predictive modeling technique that assesses the affiliation between the variables.

BC is a heterogeneous disease with great diversity in morphology and clinical behavior. The recurrence of BC after complete treatment is common; therefore, the prediction of BC recurrence is a crucial factor for successful treatment and follow-up planning. The number of invasive nodes, degree of malignancy, and irradiation were identified to be significantly associated with BC recurrence in the BC Wisconsin dataset. The nomogram developed based on this dataset can be a valuable tool in guiding appropriate treatment modalities based on the risk of recurrence. Machine learning and data mining methods can be the future of the clinical decision process. Regarding factors influencing breast cancer prognosis, screening appropriate predictors as independent variables is an important step in model construction. In previous studies, predictors mostly included patients' demographic characteristics, medical history, treatment information, and the clinicopathological characteristics of tumors at different disease stages.

In this study, demographic risk factors were also found to be individually important in our BC risk prediction model, although their importance scores were not equal. Similar to the genetic variants, combinations of demographic risk factors yielded a higher risk prediction accuracy than the individual demographic risk factors. Genomic and non-genomic entered into many predictive models as predictors, given that these factors represent breast cancer prediction.

Gender

In cancer, susceptibility is generally higher in males although some cancers are more common in women. The same is true for autoimmunity in which females have an overall higher susceptibility, but males are more susceptible for few of them. In this case, women are more prone to BC than men. In this case, the gender factor is crucial in enhancing breast cancer prediction.

Tumor Size

Tumor size was a critical clinical factor with considerable prognostic and predictive value for T1 breast cancer, and it should be selectively incorporated into the current staging system to facilitate prediction of death and recurrence risk. Tumor size, although an important predictor variable was not found to be significant in multivariate analysis in our study. This finding suggested that risk factors for recurrence vary considerably among

different study populations, and therefore, it is desirable to have careful selection criteria based on institutional data. The proposed model can achieve a higher performance on cancer tumor classification using gene expression data. Hence, the tumor stage factor is crucial in enhancing breast cancer prediction.

Age

Age also can be considered a surrogate measure for the complex biological processes associated with aging. The risk of receiving a diagnosis of different types of cancer varies throughout a person's life span. Aging increases cancer risks in our bodies in several ways. The older we are, the higher the proportion we acquire of cells with mutations. And these cells create populations of high risk for recruiting cancer-initiating cells. In this case, the age factor is crucial in enhancing breast cancer prediction.

Hereditary Factor

The hereditary factor is depicted through the fact that cancer is a genetic disease that is, cancer is caused by certain changes to genes that control the way cells function, especially how they grow and divide. Cancer-causing genetic changes can also be acquired during one's lifetime, as the result of errors that occur as cells divide or from exposure to carcinogenic substances that damage DNA. Hence, the hereditary factor is crucial in enhancing breast cancer prediction.

Strategies based on predictive genomics and cancer hallmarks for cancer biomarker identification have also been published to predict cancer risk and patient outcomes. These strategies often measure alterations in pre-defined cancer susceptibility genes. Apart from successfully generating robust cancer prognostic and diagnostic gene signatures, these strategies are often limited to a set of pre-selected candidate genes. The proposed approach in this study, which is free from pre-selection of risk factors, can be integrated with hallmark-based strategies to further enhance predictions of cancer risk and search for optimal interactions. Indeed, the present study can also be extended to other multifactorial diseases.

This study developed a regression model model of individual BC patients using the machine learning method. This model was developed using BC-related clinicopathologic factors at the time of curative surgery and consecutive clinical factors that have been

identified during the BC surveillance period. In this study, the researcher proposed an ML approach to efficiently combine genetic variants with BC risk factors related to both familial history and oestrogen metabolism and to search for optimal interactions among them. The proposed approach considerably increased the BC risk prediction accuracy compared to systems based solely on genetic variants or demographic risk factors for BC.

The results of this study should be interpreted in light of some limitations. Because the research was limited to a single institution, the results might not be generalizable to other cancer patients in other settings. Therefore, the findings should be validated using samples from other institutions to confirm generalizability. Nonetheless, our model is the first machine learning-based BC prognosis model developed using clinical information at both BC diagnosis and follow-up. Moreover, the model produced high AUC scores that remained consistent for several years after the completion of BC treatment.

5.3 Contributions of the Study

The biological theory was fundamental in the research since it facilitated the reliance advancements in both biological data generation and machine learning methodologies to conduct an analysis that led to the discovery of complex biological data. Additionally, the Machine Learning theory was essential in the research since it facilitated the creation of mathematical models that were able to highlight crucial aspects of machine learning to solve the issue of breast cancer prediction. The theory was also fundamental in proving guarantees for algorithms and developing machine learning algorithms that met the desired criteria. The theory was also important in facilitating the creation of a model that is confident in making prediction from limited data through mathematical analysis of general issues.

The empirical review was vital in guiding the research. In this case, the empirical data highlight the challenges associated with machine learning in cancer prognosis prediction to achieve high performance. In this case, the review proved that deep learning would potentially improve cancer prognosis. Additionally, it was revealed that potential gene signatures and sub-network biomarkers are biologically meaningful and can yield significantly high accuracy in predicting breast cancer outcomes after treatment. The empirical review also found that the combination of clinical data with molecular data might be the future direction for cancer prognosis and prediction (Cammarota et al, 2020). All these factors were employed in the creation of the model. The researcher proposed an ML approach

to efficiently combine genetic variants with BC risk factors related to both familial history and oestrogen metabolism and to search for optimal interactions among them.

5.3.1 Model Comparison

The created model can be compared to existing models based on its operation. In this case, Discrimination and calibration are the two main components of accuracy in a risk assessment model. Discrimination is the ability to distinguish benign abnormalities from malignant ones. Although assessing discrimination with area under the receiver operating characteristic (ROC) curve (AUC) is a popular method in medical community, it may not be optimal in assessing risk prediction models that stratify individuals into risk categories. n this setting, calibration is also an important tool for accurate risk assessment for individual patients. Calibration measures how well the probabilities generated by the risk prediction model agree with the observed probabilities in the actual population of interest. There is a trade-off between discrimination and calibration, and a model typically cannot be perfect in both. In general, risk prediction models need good discrimination if their aim is to separate malignant findings from benign ones, and good calibration if their aim is to stratify individuals into higher or lower risk categories to aid in decision-making and communication.

ANN can accurately estimate the risk of breast cancer using a dataset containing demographic data and prospectively-collected mammographic findings. However, the proposed model is unique based on its reliance on genomic and non-genomic factors that enhance the prediction of BC. Furthermore, contrary to previously developed CADx models in breast cancer risk prediction, the expand the evaluation of models beyond discrimination by measuring the accuracy of the estimated probabilities themselves using calibration metrics.

5.4 Recommendations for Future Research

The variables were vital in decision-making analysis in relation to breast cancer. In the future, the possible mechanisms underlying the occurrence and development of breast cancer could be further studied from these perspectives, which also suggests that more suitable predictors for clinical practice can be identified. The ML predictive model applied in this research can be translated into tools for clinical treatment decision-making. Visualization

of some of the outcomes will be implemented in the research database and used by the clinicians at the hospital to enhance the prediction of breast cancer patients.

Following accurate comparison between the model, it was established that Support Vector Machine achieved a higher efficiency of 97.2%, Precision of 97.5%, AUC of 96.6% and outperforms all other algorithms. In conclusion, Support Vector Machine has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy and precision. It should be noted that all the results obtained are related just to the KNH database, it can be considered as a limitation. Therefore, it is necessary to reflect for future works to apply these same algorithms and methods on other databases to confirm the results obtained via this database, as well as, in our future works, we plan to apply our and other machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

The research was limited to a single institution. Additionally, it was limited to BC prediction. Therefore, the results might not be generalizable to other cancer types. According to the research, the choice of the most appropriate algorithm depends on many parameters including the types of data collected, the size of the data samples, the time limitations as well as the type of prediction outcomes. Therefore, the future of cancer modelling new methods should be studied for overcoming data sample size, and type of prediction outcome limitations. A better statistical analysis of the heterogeneous datasets used would provide more accurate results and would give reasoning to disease outcomes. Further research is required based on the construction of more public databases that would collect valid cancer dataset of all patients that have been diagnosed with the disease. Their exploitation by the researchers would facilitate their modelling studies resulting in more valid results and integrated clinical decision making.

References

Babaian, R. J., Fritsche, H., Ayala, A., Bhadkamkar, V., Johnston, D. A., Naccarato, W., & Zhang, Z. (2000). Performance of a neural network in detecting prostate cancer in the prostate-specific antigen reflex range of 2.5 to 4.0 ng/mL. *Urology*, *56*(6), 1000-1006.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J., ... & Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature reviews gastroenterology & hepatology*, *17*(10), 635-648.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., ... & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), 045002.

Das, A., Rad, P., Choo, K. K. R., Nouhi, B., Lish, J., & Martel, J. (2019). Distributed machine learning cloud teleophthalmology IoT for predicting AMD disease progression. *Future Generation Computer Systems*, *93*, 486-498.

Davi, C., & Acioli-Santos, B. (2019). Severe dengue prognosis using human genome data and machine learning. *IEEE Transactions on Biomedical Engineering*, *66*(10), 2861-2868.

Doria-Rose, N., Suthar, M. S., Makowski, M., O'Connell, S., McDermott, A. B., Flach, B., ... & Kunwar, P. (2021). Antibody persistence through 6 months after the second dose of mRNA-1273 vaccine for Covid-19. *New England Journal of Medicine*, *384*(23), 2259-2261.

Du, K. L., & Swamy, M. N. S. (2019). Elements of computational learning theory. In *Neural networks and statistical learning* (pp. 65-79). Springer, London.

Feld, S. I. et al. (2018). Improving breast cancer risk prediction by using demographic risk factors, abnormality features on mammograms and genetic variants. *In AMIA Annual Symposium Proceedings*, 1253–1262

Guertler, M. R., Kriz, A., & Sick, N. (2020). Encouraging and enabling action research in innovation management. *R&D Management*, *50*(3), 380-395.

Hajiloo, M. & Damaraju, S. (2013). ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. *BMC bioinformatics*, *14*(1), 1-15.

Harks, T., & Klimm, M. Algorithmic Game Theory LNCS 12283.

Hill OT, Mason TJ, Schwartz SW, et al. Improving prostate cancer detection in veterans through the development of a clinical decision rule for prostate biopsy. BMC Urol 2013;13:6. doi:10.1186/1471-2490-13-6

Hosseini, M. P., Lu, S., Kamaraj, K., Slowikowski, A., & Venkatesh, H. C. (2020). Deep learning architectures. In *Deep learning: concepts and architectures* (pp. 1-24). Springer, Cham.

Ikehi, M. E., Onu, F. M., Ifeanyieze, F. O., Paradang, P. S., Nwakpadolu, M. G., Ekenta, L. U., & Nwankwo, C. U. (2019). Survey on Sample Sizes of Postgraduate Theses in Agricultural Education and Extension in Universities of Nigeria. *Journal of Extension Education*, *31*(1).

Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes of seismic noise signals using unsupervised learning. *Geophysical Research Letters*, *47*(15), e2020GL088353.

Kundrod, K. A., Smith, C. A., & Richards-Kortum, R. (2019). Advances in technologies for cervical cancer detection in low-resource settings. *Expert review of molecular diagnostics*, *19*(8), 695-714.

Learned, K., Durbin, A., Currie, R., Kephart, E. T., Beale, H. C., Sanders, L. M., ... & Bjork, I. M. (2019). Barriers to accessing public cancer genomic data. *Scientific data*, *6*(1), 1-7.

Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., & Antoniou, C. (2019). BOADICEA: a comprehensive breast cancer risk prediction model

incorporating genetic and nongenetic risk factors. *Genetics in Medicine*, 21(8), 1708-1718.

Libes, J. M., Seeley, E. H., & Kenyan Wilms Tumor Consortium. (2014). Race disparities in peptide profiles of North American and Kenyan Wilms tumor specimens. *Journal of the American College of Surgeons*, *218*(4), 707-720.

Louro, J., Román, M., Posso, M., Vázquez, I., Saladié, F., Rodriguez-Arana, A & BELE and IRIS Study Groups. (2021). Developing and validating an individualized breast cancer risk prediction model for women attending breast cancer screening. PloS one, 16(3), e0248930.

Mavaddat, N., Rebbeck, T. R., Lakhani, S. R., Easton, D. F., & Antoniou, A. C. (2010). Incorporating tumour pathology information into breast cancer risk prediction algorithms. *Breast Cancer Research*, *12*(3), 1-12.

Mitchell, M. (2019). Selecting the Correct Predictive Modeling Technique. https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine learning: algorithms and applications*. Crc Press

Odeyemi, O. (2020). *Integrated Machine Learning and Bioinformatics Approaches for Prediction of Cancer-Driving Gene Mutations* (Doctoral dissertation, Chapman University).

Omar Ali, N. (2020). A Comparative study of cancer detection models using deep learning.

Osewe, E. J., & Muturi, W. Effects of e-banking innovations on the financial

Otieno, E.S. (2008). The Pattern of Breast Diseases at Kenyatta National Hospital. *The Analysis of African Surgery*, 2, 25-29.

Phan, L. T., Nguyen, T. V., Luong, Q. C., Nguyen, T. V., Nguyen, H. T., Le, H. Q., ... & Pham, Q. D. (2020). Importation and human-to-human transmission of a novel coronavirus in Vietnam. *New England Journal of Medicine*, *382*(9), 872-874.

Salem, H., Soria, D., Lund, J. N., & Awwad, A. (2021). A systematic review of the applications of Expert Systems (ES) and machine learning (ML) in clinical urology. *BMC Medical Informatics and Decision Making*, *21*(1), 1-36.

Shastry, K. A., & Sanjay, H. A. (2020). Machine learning for bioinformatics. In *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications* (pp. 25-39). Springer, Singapore.

Tresp, V., & Yu, S. (2016). Going digital: a survey on digitalization and large-scale data analytics in healthcare. *Proceedings of the IEEE*, *104*(11), 2180-2206.

Tsai, T.-Y., You, J.-F., Hsu, Y.-J., Jhuang, J.-R., Chern, Y.-J., Hung, H.-Y., Yeh, C.-Y., et al. (2021). A Prediction Model for Metachronous Peritoneal Carcinomatosis in Patients with Stage T4 Colon Cancer after Curative Resection. Cancers, 13(11), 2808. MDPI AG. Retrieved from http://dx.doi.org/10.3390/cancers13112808

Waks, A. G., & Winer, E. P. (2019). Breast cancer treatment: a review. *Jama*, *321*(3), 288-300.

Wang, F., Zhao, N., Gao, G., Deng, H. B., Wang, Z. H., Deng, L. L., ... & Lu, C. (2020). Prognostic value of TP53 co-mutation status combined with EGFR mutation in patients with lung adenocarcinoma. *Journal of Cancer Research and Clinical Oncology*, *146*(11), 2851-2859.

Wanjawa, D. S., Yugi, C. T., & Muli, W. M. Contribution of Agricultural Loans Accessibility to Performance of Small Holder Sugar Cane Farmers in Kakamega County, Kenya.

Zairis, S. (2018). *Quantitative Approaches to the Genomics of Clonal Evolution* (Doctoral dissertation, Columbia University).

Zhu, W., Xie, L., Han, J., & Guo, X. (2020). The application of deep learning in cancer prognosis prediction. *Cancers*, *12*(3), 603.

Zhu, W., Xie, L., Han, J., & Guo, X. (2020). The application of deep learning in cancer prognosis prediction. *Cancers*, *12*(3), 603.

Zitnik, M. et al. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50, 71–91.

APPENDICES

Appendix 1: Research schedule

Activity	May-June	Jul-Aug	Sep-Oct
Concept paper development			
Proposal -Chapter one			
Literature review and Methodology			
Questionnaire formulation			
Field Data Collection			
Data Analysis			
Report Writing			
Submission			

Appendix ii: Resources and Budget

No.	Items	Cost in KSHS.
1	Stationery, typing papers, pens, flash disk	10,000.00
2	Secretarial services	20,000.00
3	Printing	5,000.00
4	Binding	6,000.00
5	Mobile phones expenses	6,000
6	Communication and telephone Services	10,000.00
	TOTAL	**57,000.00**