MACHINE LEARNING BASED CERVICAL CANCER DETECTION MODEL IN WESTERN KENYA

JOHN FLAVIAN MURERE

A THESIS SUBMITTED TO THE SCHOOL OF SCIENCE IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE CONFERMENT OF THE DEGREE OF MASTER OF SCIENCE IN MATHEMATICS (BIOSTATISTICS) OF THE UNIVERSITY OF ELDORET, KENYA

2025

DECLARATION

Declaration by the Candidate

This thesis is my original work and has never been presented for the award of an academic degree in any other university and should not be copied, or reproduced in any format without written authority from the author and/or University of Eldoret.

John Flavian Murere

_____ Date_____

SSCI/MAT/M/011/22

Approval by the Supervisors

This thesis is submitted with our approval as the university supervisors.

_____ Date_____

Dr. Julius Koech, PhD

School of Science

Department of Mathematics and Computer Science

University of Eldoret, Kenya

_____ Date_____

Dr. Samson Wangila, PhD

School of Science

Department of Mathematics and Computer Science

Pwani University, Kenya

DEDICATION

I dedicate this thesis to my mother Mrs. Emily Juma Murere and my brothers Mark Dawson, Bravin Wanyonyi and Victor Junior.

ACKNOLEDGEMENT

I would like to express my sincere gratitude to Dr. Lubao Wanyonyi Murere for his invaluable guidance throughout the research process. He guided me from proposal development to thesis writing and correction phases. His encouragement contributed greatly to the successful completion of this work. I also express my heartfelt appreciation to Mitchelle Ochieng for their invaluable assistance in collecting and organizing the data for this study.

I am also thankful to Mr. Titus Kigen for his technical assistance with some parts of the data analysis. His expertise and patience were instrumental. He helped ensure that the results were accurate and of high quality.

Lastly, I acknowledge all the students and staff at the Mathematics and Computer Science Department, University of Eldoret, who, in one way or another, supported and encouraged me throughout this journey. Above all, I give thanks to the almighty God for granting me good health, strength, and determination to accomplish this work.

ABSTRACT

Cervical cancer is the leading cause of cancer-related deaths among Kenyan women, with approximately 3,200 deaths reported annually, driven mainly by low screening uptake (16%) and late diagnosis. The aim of this study was to develop a machine learning based model that would enhance the detection of cervical cancer in Western Kenya, a region that has limited healthcare resources. This study used a cross-sectional study design where data from 968 women were collected, including information on demographics, reproduction, and clinical characteristics. Data was collected from health facilities. The study showed that 93.7% (n = 907) had no biopsy-confirmed abnormalities, while 6.3% (n = 61) had abnormalities. There were five machine learning models (Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, and Artificial Neural Network) that were trained on 70% of the data (training set) and tested on 30% of the data (testing set). The random forest model achieved the highest accuracy (94.33%) and specificity (98.37%), which outperformed the other models and traditional methods like Human papilloma virus (HPV) testing (70-80% specificity) and Pap smear (>90% specificity) for confirming negative cancer cases. The logistic regression model had the highest sensitivity of 70% which was comparable to the Pap-smear method (60-95% sensitivity), but it was lower than the HPV testing, with a sensitivity greater than 90% which makes it suitable for initial cervical cancer screening. The Pap smear results and use of hormonal contraceptives emerged as the key significant predictors of cervical cancer, which supports targeted screening strategies. The findings from this study confirmed there was a significant difference in model performance with partial superiority over existing methods and the influence of key cervical cancer risk factors. The combined approach of using a random forest model for confirmation and logistic regression for screening could optimize cervical cancer screening further in the resource-constrained Western setting. This study has underscored the potential that machine learning has in addressing cervical cancer disparities in Western Kenya, with implications for both public and private health interventions and future research work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS, ACRONYMS, AND SYMBOLS

ANN: Artificial Neural Network

CC: Cervical cancer

CCDCM: Cervical cancer Detection and Classification Model

DL: Deep Learning

GBM: Gradient boosting model

GV: Gini Value

HPV: Human Papilloma Virus

KNN: K-Nearest Neighbors

LM: Logit model

ML: Machine Learning

MLA: Machine Learning Algorithm

MoH: Ministry of Health

MR: Mortality Rate

RFM: Random forest model

SM: Screening Methodologies

SVM: Support vector machines

PID: Pelvic Inflammation Disease

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Cervical cancer is one of the most common cancers among females and the leading cause of mortality in many developing countries (Arbyn et al., 2020). Cervical cancer remains a global health problem and is the fourth most common cancer affecting women worldwide (Lilhore et al., 2022). Over time, a significant gap persists in developing countries where approximately 90% of the global burden of mortality occurs. The mortality rate from cervical cancer in developing countries is approximately 18 times greater than in developed countries (Akinyemiju, 2012). Human life is faced with challenges that often arise unexpectedly. Among the many challenges women encounter throughout their lives, cervical cancer stands out as one of the most serious conditions they may face (Martin et al., 2009). The high number of deaths attributed to cancer has been linked to the late diagnosis (stage 3 or 4) of the disease, making it hard for health practitioners to treat or manage the disease (Sfeir et al., 2018). Cancer is a medical condition characterized by an uncontrolled growth of certain cells in the body, which can spread to other parts of the body (National Cancer Institute, 2021). Cancer is a medical condition characterized by symptoms such as unexplained weight loss, changes in skin color, persistent severe pain, a chronic cough, Swelling or lumps in the lymph nodes, and consistent headaches, among others (Cancer Research UK, 2015). Cervical cancer primarily impacts the cervix which is the lower portion of the uterus that links to the upper part of the vagina (Berek & Berek, 2020). The signs and symptoms of cancer vary based on its stage, type, and the affected

area. Thus, for cervical cancer, the main symptoms include vaginal bleeding, unusual vaginal discharge, pelvic discomfort, bleeding after bowel movements, blood in urine, and weight loss.

The leading contributing factor to the high mortality rate of cervical cancer among women is a lack of awareness of the importance of early detection (Purnami et al., 2016). The signs and symptoms of cervical cancer are relatively hard to detect at an early stage, leading to late detection in most women when the signs start to develop (Yang et al., 2018). Middle and low-income countries account for 83% of the world's cervical cancer burden, yet their screening coverage is only 19%. In contrast, high-income countries, which tend to have lower cervical cancer burden, enjoy a screening coverage of 63% (Matenge & Mash, 2018). In developed countries such as the United States and the United Kingdom, the mortality rate associated with women diagnosed with cervical cancer ranges from 40 to 42% (LaVigne et al., 2017), which is almost double the mortality rate in Africa and South Asian countries (78%), which are less developed (Torre et al., 2015). Cervical cancer is the second most common in terms of incidence and the leading cause of cancer-related deaths among women in Africa, responsible for almost 15% of such deaths (Momenimovahed et al. 2023; World Health Organization, 2020).

In Africa, cervical cancer is the most prevalent cancer among women aged between 14 and 70. Likewise, approximately 570,000 women are diagnosed with cervical cancer each year, and about 284,000 die from the disease (World Health Organization, 2022). Cervical cancer accounts for 11% of all cancer diagnoses among women and 23% of cancer-related deaths among African women. Approximately 90% of deaths from cervical cancer occur

in low and middle-income countries, highlighting significant disparities in cervical cancer outcomes stratified across different socio-economic contexts (Arbyn et al., 2020).

A key risk factor for cervical cancer is the long-term infection with high-risk HPV strains, which are viruses spread through sexual contact. Notably, Human Papilloma Virus (HPV) types 16 and 18 are responsible for approximately 70% of all cervical cancer cases worldwide (Walboomers et al., 1999).

In addition to the persistent HPV infection, the other potential risk factors for cervical cancer include smoking, having a weakened immune system, and use of hormonal contraceptives (Kashyap et al., 2019). A full-term pregnancy at a young age, having multiple sexual partners, or being in a relationship with a partner who has multiple sexual partners increases the risk of cervical cancer (Plummer et al., 2016).

1.2 Statement of the problem

Cervical cancer continues to be a major health problem, particularly in low-income regions worldwide. It is projected that deaths from cervical cancer will increase by 4,430 in 2030 and 6,200 in 2040 in Kenya (MoH, 2022). Despite cervical cancer being one of the most easily preventable cancers through screening successfully implemented in high-income regions like the United States and the United Kingdom, the number of diagnosed cases and deaths continues to rise in Kenya. The disparity in cancer statistics between high-income and low-income regions underscores the urgent need for enhanced healthcare screening methods to alleviate the burden of this disease.

In Sub-Saharan Africa, cervical cancer mortality rates are among the highest globally, with an overall rate of about 34.5 per 100,000 women (Burt et al., 2021). In Eastern Africa,

the rate is notably elevated but lower than the Sub-Saharan average at 25.3 per 100,000 (Tadesse, 2015). The rising incidence of cervical cancer in developing countries is attributed to the inadequate population-wide screening programs, the low awareness about cervical cancer, the unequal access to healthcare services, and the widespread poverty (Al-Naggar, 2022; Jedy-Agba et al., 2020).

The prevention of cervical cancer through screening strategies remains a crucial tool for eradication, particularly in developing countries where HPV vaccination is often less convenient, less accessible, and less accepted (Petersen et al., 2022). Though the governments have made significant strides in increasing screening uptake, however, the number of deaths and cases continues to rise, with projections from the Ministry of Health indicating this trend will persist. Factors contributing to this disparity include access to detection, screening, and other healthcare services. Research indicates that numerous potential risk factors impact cervical cancer among women, which creates challenges in predicting treatment outcomes

This study aimed to minimize these existing problems by developing a machine learning-based algorithm for cervical cancer detection in Western Kenya that is individually based. This study aimed to develop a reliable detection model that is specific to Western Kenya, which is a low-income region that would be used to accurately predict an individual's likelihood to develop cancer based on their unique characteristics, to bridge the gap in cervical cancer mortality rate between low income and high-income regions.

1.3 Justification of the study

The severity of cervical cancer is evident in Western Kenya, where it is the leading cause of cancer-related deaths. This underscores the urgent need for targeted interventions. This

study aims to leverage machine learning tools to develop a prognostic model that will assist clinicians, doctors, and healthcare practitioners in the early detection of cervical cancer in assessing the risk of patients developing the disease.

Machine learning algorithms offer the potential to unearth complex relationships and patterns within the individual's data, which will aid in targeted screening by identifying individuals at a high risk of cervical cancer. With early detection and targeted screening, the model can help reduce the burden of cancer on healthcare facilities.

The justification of this study lies in leveraging machine learning techniques to develop a tool that will empower healthcare workers to make more informed decisions that will lead to more timely intervention by using clinical information, medical history, and demographic characteristics.

This research aims to use the developed tool to lower the mortality rate and cases associated with cervical cancer by enhancing accuracy and providing timely results. This research is interested in providing technological solutions and offering a positive transformation in the healthcare sector by addressing the inequalities between low and high-income regions.

1.4 General objective

To develop and evaluate a Machine Learning-Based Model for cervical Cancer detection in Western Kenya.

1.4.1 Specific objectives

i. To develop machine learning models for the detection of cervical cancer among women in Western Kenya, and to evaluate their performance and consistency across unseen data.

ii. To compare the predictive capability of the developed machine learning model with existing cervical screening approaches (HPV testing, Pap smears and cytology) in Western Kenya.

iii. To determine the most influential risk factor associated with cervical cancer screening among women in Western Kenya.

1.5 Hypothesis

Ha1: There is a statistically significant difference among the machine learning fitted models to detect cervical cancer in terms of accuracy, sensitivity, and specificity on unseen data.

Ha2: The developed machine learning model will be significantly better in detecting and classifying cervical cancer disease in terms of predictive ability than existing screening methods among women in Western Kenya.

Ha3: There is at least one influential risk factor that influences the cervical cancer screening among women in Western Kenya.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This section documents the related materials used under cervical cancer screening in both developed and developing countries. Also this chapter outlines the ML methods that have been used in the prediction of health outcomes among women with cancer-related cases. History of cervical cancer globally and in Kenya which is one of the developing countries is presented below.

2.2. Cervical Cancer in Developed Countries

Worldwide, cervical cancer is the second most common disease among women of reproductive age. It has been identified as the leading cause of incidence and mortality in many countries (Wu et al, 2022).

2.2.1 The Epidemiology and Burden of Cervical Cancer Globally

Incidence and mortality rates for cervical cancer have significantly declined in most developed countries, thanks to widespread screening and early detection efforts. However, it remains a public health concern in certain populations. In the U.S., the American Cancer Society reported approximately 13,820 new cases and 4,360 deaths from cervical cancer in 2024 (Siegel et al., 2024). Countries with well-established screening programs, such as the UK and Australia, have experienced reductions of over 70% in cervical cancer mortality over the past 40 years (Arbyn et al., 2020).

2.2.2. Screening Programs

Regular screening with **Pap smears** and **HPV testing** has led to early detection and significant reductions in cervical cancer incidence and mortality in developed countries. In countries like Sweden and the Netherlands, organized population-based screening programs have achieved coverage levels above 80%. The U.S. uses a combination of Pap smear and HPV testing for women aged 30 to 65 years, with extended screening intervals.

2.2.3. HPV Vaccination Programs

The introduction of the HPV vaccine in many developed countries has further reduced cervical cancer rates, especially among younger women. Australia is on track to eliminate cervical cancer as a public health issue by 2035 due to high vaccine uptake and screening coverage (Hall et al., 2019). The UK and the US have integrated HPV vaccination into national immunization schedules, targeting adolescents.

2.2.4. Disparities and Access

Despite overall progress, disparities in cervical cancer outcomes persist in developed countries due to socio-economic, racial, and geographic factors. In the U.S., Black and Hispanic women are more likely to be diagnosed at a later stage and have higher mortality rates (Cohen et al., 2023). Rural populations and uninsured women face barriers to both screening and treatment ().

2.2.5. Innovations in Treatment and Management

Developed countries continue to lead in innovative treatments, including targeted therapy, immunotherapy, and robotic surgery.

- The use of immunotherapeutic agents such as checkpoint inhibitors has shown promise for recurrent or advanced cervical cancer (Han et al., 2022).
- Personalized treatment approaches based on molecular profiling are becoming increasingly common (Wang & Wang, 2023).

2.3 Cervical Cancer in Developing Countries

Cervical cancer remains one of the most significant public health challenge in developing countries where it is the leading cause of cancer related deaths among women due to limited access to treatment and screening services (Arbyn et al., 2020). Approximately 85% of the global cervical cancer cases were reported in the low and middle income countries with the Sub-Saharan African countries and the Southern Asian countries carrying the highest incidence burden (Arbyn et al., 2020).

The lack of infrastructure, inadequate healthcare systems and low awareness about cervical cancer contributes to the late diagnoses which reduces the survival rates in developing countries (Rayner et al., 2023). The regular cervical cancer screening and human papillomavirus (HPV) could prevent most cancer cases but the distribution and cost in developing countries hinders the implementation (Sankaranarayanan et al., 2019). The strengthening of the healthcare systems through innovation and international support are critical for reducing disparity of cervical cancer cases between developed and developing countries (Denny et al., 2022).

2.4 Cervical Cancer in Kenya

Cancer is ranked as the third leading cause of death of cardiovascular and infectious diseases in Kenya (World Health Organization, 2020). It is the leading cause of cancer

related deaths in Kenya, with approximately 3,200 deaths reported in 2020 (Bruni et al., 2018).

The uptake of cancer screening services has been low, with only 16% of women in Kenya having gone for the screening (Nwabichie et al., 2017). Nwabichie and others indicated that lack of awareness, stigma, and access to healthcare were significant factors that hindered the high uptake of screening services. The world health organization projects the number of cervical cancer cases to be 99,000 and deaths to be approximately 4,430 by 2030 (World Health Organization, 2022).

The higher number of deaths associated with cancer is the lack of access to cancer screening facilities, social demographic factors such as income levels, early sexual debut, and high rates of human papillomavirus infections (World Health Organization, 2022). The prevalence of human papillomavirus is estimated to be 20% in Kenya (Bruni et al., 2018), while 38.6% of Kenyans are poor line which means they cannot afford proper healthcare without compromising on either to have food or shelter (Kenya National Bureau of Statistics, 2023).

2.5 ML in Healthcare

The emergence of machine learning in healthcare sector has been transformative due to its unrivalled capabilities for prediction, analysis and decision making (Shailaja, Seetharamulu, & Jabbar, 2018). The application of machine learning spans over a wide spectrum of the healthcare domains, revolutionizing how healthcare is delivered and practiced and optimizing clinical decisions with the sole purpose of improving the patient's wellbeing (Siddique & Chow, 2021). The major area where machine learning has been used in the healthcare sector include:

2.5.1 Disease Diagnosis

Machine learning can processes patients' medical data, medical imaging, genetic information and medical history to develop an algorithm to assist in accurate and timely diagnosis of the disease.

A notable example of disease diagnosis using machine learning is in the study conducted by Oh et al. (2021) that used deep learning and ultra-wide-field fundus images in the early detection of retinopathy in retinal images. These researchers developed a diabetic retinopathy detection system based on fundus photography and deep learning to diagnose diabetic retinopathy. The developed model outperformed the optic disc and macula centered image in a statistical sense in the early detection of the disease using 7-standard field image extracted from ultra-wide-field fundus photography. The experiment also showed that the deep learning fitted algorithm was efficient and saved time and workforce compared to other existing diagnostic methods.

2.5.2 Drug discovery

Machine learning is used to accelerate the identification of potential drugs by identifying molecular patterns, predicting the interactions in drug receptor, and analyzing biological data (Duch et al., 2007).

2.5.3 Personalized treatment plan

Machine learning in personalized treatment refers to the use of machine learning algorithm to identify patterns that will help make personalized treatment recommendation by analyzing large amounts of data (Pitts et al., 2009). In the application of data-intensive biomedical technology in research, it is seen that human beings differ greatly at the

exposure, physiological, biochemically, genetically, behaviorally, their response to treatment, and in regard to disease processes (Schork, 2019). Schork asserted that these human differences demands a need to have a personalized medicine routine or medicines which are unique based on the features possessed by an individual patient. The personalized treatment approach ensures that treatment is tailored towards maximum efficacy and minimized adverse effect (Ahmed et al., 2020)

2.5.4 Healthcare Management

Machine learning supports healthcare management system by optimizing resource allocation, predicting patient admission numbers, improvement of hospital operation and delivery of timely patient's results. Creswell and Sheikh (2013) indicated that these application reduced the cost of operation of a healthcare facility, cost of healthcare for the patient and enhanced the patient's outcome.

2.6 Cervical cancer testing instruments and their accuracy

2.6.1 Pap smear:

The Pap smear, a cornerstone in cervical cancer screening, has demonstrated consistent effectiveness in identifying abnormalities (Bora et al., 2017). Studies report an average sensitivity ranging from 60% to 95%, while specificity is generally high, exceeding 90% (Song et al., 2017). However, challenges such as inter-observer variability and false negatives highlight the need for continuous improvement.

2.6.2 HPV Testing:

Human Papillomavirus (HPV) testing has emerged as a robust primary screening tool due to its strong association with cervical cancer (Haile, 2019). High-risk HPV types contribute

significantly to carcinogenesis. HPV testing exhibits high sensitivity (usually above 90%) and specificity (typically around 70-80%), making it a valuable component of cervical cancer screening programs (Haile, 2019).

2.6.3 Liquid-Based Cytology:

Liquid-based cytology, an evolution of conventional Pap smears, offers improved sample preparation (Kumaresan, 2021). Studies indicate comparable sensitivity to Pap smears, ranging from 70% to 90%, with a potential for increased specificity (Masenya, 2011). The reduction in inadequate specimens enhances overall accuracy.

2.7 Theoretical Framework

This section lays down the theoretical foundation for this study, including fundamental machine learning concepts, their potential application, and their incorporation into medical classification within western Kenya.

2.7.1 Core Principle

Machine learning uses algorithm to give a computer the ability to learn from the data, identify patterns and make predictions based on the data without explicit programing (Jordan & Mitchell, 2015). Machine learning operates under ideas and principles that underpin its functionality:

2.7.2 Supervised Learning

This is one of the fundamental concept of machine learning where the algorithm learn from the data that is labelled. The dependent variable and the independent variables are known. The objective of supervised learning algorithm is to learn a mapping function

developed from the training data that can be used to predict unseen data (Cunningham et al., 2008). Some of the supervised machine learning model are:

2.7.2.1 Artificial Neural Network

Artificial neural network (ANN) is one of the most common and fundamental method of deep learning which is a subfield of machine learning. NN are brain inspired system created to mimic how the human brain works (Jain, Mao, & Mohiuddin, 1996)

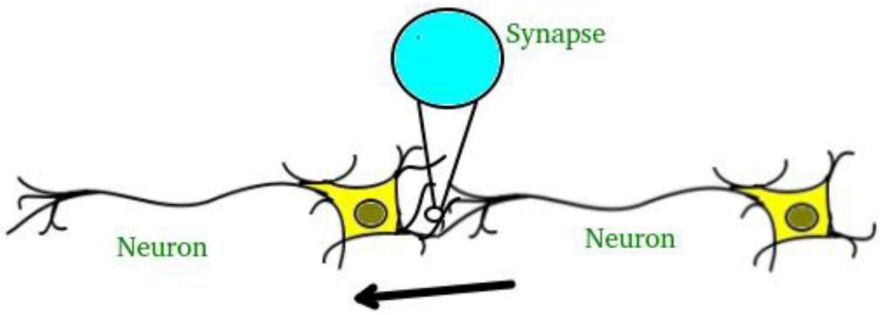

Figure 1: Artificial Neurons (Source: Wu et al, 2022)

The ANN is made up of interconnected units or nodes (artificial neurons) that nearly resemble the neurons in the human brain. In human being, the human central nervous system contains the neurons which are cells. The connecting region between axioms and dendrites is called synapses. Each link allows the transmission of synapses with weighted data acquisition and mathematical computations. Psychologist Frank Rosenblatt was the first person to create an artificial neural network based on this paradigm (Sharma, 2017). The ability of artificial neural network to learn non-linear complex relation automatically from the data has made it a method of choice in a healthcare sector and other different sectors (He, Zhang, Ren, & Sun, 2016). The technological advancement which has been characterized by advanced computing power has made it easy to train and test artificial neural network for state of the art results classifications (Mahanama, 2020).

Artificial neural network calculates the function of the input by propagating the computed values from the input neurons to the output neuron using weight of the parameters in the data. Just as the external stimuli are required for biological beings to learn, the input and output pairing which is the training metrics for artificial neural network works also serves as the external stimuli. For example, in this study, the data might contain clinical, medical history and other patients data and their output, the output are diagnosed and not diagnosed with cancer. The training data pair is supplied into the neural network which will make prediction about the label of the output.

2.7.2.2 Random forest

Random forest, a robust and versatile machine learning modelling technique has also widely been used in different sectors such as health. Random forest functions on district rules which are different from the neural network which is modelled after how a human brain works. Breiman (2001) introduced the concept of random forest.

Random forest is an ensemble learning method which relies on more than one decision tree as it leverages the strength of various decision tree to reach at the best robust model. The forest is made up of many decision tree where each decision tree is constructed based on different data subset of the data the models are being trained on and the trees work together to produce one model by either voting on the best or averaging (Breiman, 2001).

Breiman (2001) indicated that random forest thrive when dealing with complex dataset and nonlinear data. They are more suited for tasks that involve high-dimensional data with a lot of noise as they can automatically capture intricate patterns, dependencies and connections. Due to their robustness and preciseness they are popularly used in many

sector, including healthcare which requires precise classification and accurate predictions (Cutler et al., 2007)

2.7.2.3 Support Vector Machine

Support Vector Machines (SVM) constitute a prominent machine learning algorithm that has been widely employed in various domains, including medical diagnostics and disease classification. This versatile tool, introduced by Vapnik and Cortes in 1995, has gained significant attention due to its exceptional performance in binary and multiclass classification tasks (Vapnik & Cortes, 1995).

SVM operates on the principle of constructing an optimal hyperplane, or decision boundary that maximizes the margin between different classes of data points. In the context of cervical cancer detection and classification, SVM can effectively discriminate between normal and abnormal cervical cell samples. This capability is crucial for early diagnosis and intervention, particularly in regions with limited healthcare resources like Western Kenya (Arbyn et al., 2015).

One of the distinguishing features of SVM is its ability to handle both linearly and nonlinearly separable data. While a linear SVM aims to find the best-fitting hyperplane in the feature space, non-linear SVMs employ kernel functions to map the data into a higher-dimensional space where separation becomes feasible. This flexibility makes SVM suitable for capturing intricate patterns and complex relationships within cervical cancer datasets (Schölkopf & Smola, 2002).

The effectiveness of SVM in medical diagnosis, including cervical cancer detection, can be attributed to several factors. Firstly, SVM is known for its capacity to generalize well, even when dealing with high-dimensional feature spaces. In cervical cancer detection

models, these features could include clinical data, imaging results, and patient demographics. SVM excels at learning from such complex and diverse information (Chang & Lin, 2011).

SVM's margin-based approach inherently reduces the risk of overfitting, a common concern in machine learning. This property is critical when developing accurate and reliable diagnostic models, as overfitting can lead to poor generalization to new, unseen data. The ability to handle limited datasets, often encountered in medical research, is another advantage of SVM (Guyon et al., 2002).

2.7.2.5 Decision Tree

Decision Tree is a powerful machine learning algorithm that splits data into subsets recursively based on feature values to make predictions (Quinlan, 1986). It mimics human decision-making by creating a tree-like structure where each internal node represents a feature, branches denote decision rules, and leaves indicate outcomes. In healthcare, Decision Trees excel at interpretable classification, such as predicting disease risk from clinical variables (Podgorelec et al., 2002). Their transparency allows clinicians to trace prediction paths, enhancing trust and adoption in diagnostic settings.

2.7.2.5 Logistic Regression

Logistic Regression (Logit Model) is a statistical method for binary classification that models the probability of an outcome using a logistic function (Hosmer & Lemeshow, 2000). It estimates the relationship between independent variables and a dichotomous dependent variable via maximum likelihood estimation. Widely used in medical diagnostics, it provides interpretable odds ratios and performs well with linear relationships

and small datasets (Bagley et al., 2001). In cancer prediction, it effectively assesses risk factors while maintaining simplicity and robustness.

2.7.6 Unsupervised Learning

This is one of the fundamental concept of machine learning where the algorithm identify structures, patterns, or clusters within the data without explicit guidance using unlabeled data (Barlow 1989). Unsupervised can be used in healthcare industry to find hidden patients subgroups, disease prototype, and group medical information into particular area of interest (Raza & Singh, 2021).

CHAPTER THREE

MATERIAL AND METHODS

3.1 Introduction

This section outlines the methods used to achieve the specific objectives stated in Chapter One. It covers the following topics: the methods employed, the development of the research workflow model, sample size estimation, data management, mechanisms for addressing data imbalance, model feature selection, the logistic regression model, machine learning models such as random forest, support vector machines, k-nearest neighbors, and artificial neural networks, as well as model testing and evaluation, model metrics, and ethical considerations for the research project.

3.2 The Study Design

This study employed a retrospective study design where existing records were retrieved from medical patient records. The study retrospectively gathered information on the past exposures of the respondents involved. The thesis was conducted in three parts. The first phase involved understanding the problem domain, including the general and specific objectives presented in the introduction section. Information from doctors, clinicians, and other healthcare providers was used to identify eligible respondents for participation in the study.

3.3 The study site

The study site was Kakamega referral hospital and Moi teaching referral hospital in western Kenya. The two study sites were chosen given their good record keeping and the availability of the cervical cancer data among women of the reproductive age-group.

3.4 Data Collection Process

This phase involved data collection and preparation. Both close and open ended questionnaires were used for acquiring the needed respondent information the questionnaire. Thus the data was obtained from patients who had undergone cervical cancer testing and received their results. The other information was obtained from the relevant health records that are available in the health facility.

Inclusion criteria

- Biologically female participants
- Women aged 18–65 years
- Participants who are willing and able to provide informed consent

3.5 Sample Size Estimation

An appropriate sample was calculated using Fisher's formulae.

Formulae:

$$n = Z^2 * \frac{([p * (1 - p)])}{d^2} \quad \text{i}$$

Where:

Z = Critical value associated with the level of significance (1.96)

p = the estimated prevalence

d =degree of precision

Since the prevalence of CC in western Kenya is unknown, a proportion of 50% will be used

$$n = 1.96^2 * \frac{([0.5 * (1 - 0.5)])}{0.05^2} = 384.16$$

n = approximately 385 (Not less than 385)

3.6 Data Management Process and analysis

Data management process was performed using both Excel and R statistical packages. Before conducting data analysis, the data was checked for inconsistencies and outliers to ensure its accuracy and reliability. Data merging was performed to enable comparisons between the two study sites. This validation process utilized exploratory data analysis techniques, including frequency distribution tables, box plots, and descriptive statistics. Likewise, data analysis was performed using both exploratory and inferential statistics. For exploration was done using:

3.6.1 Dataset preparation

Data collection was the most important step in the development of the machine learning models as it was used to train and test the model. The social demographic characteristics, Medical history, Clinical information, quality of life, and other factors such as smoking status, alcohol consumption, and BMI collected were used as the input information in the models that were fitted. Participants who were 21 years to 70 years or sexually active, residents of Western Kenya, and individual who have been tested and have their cervical cancer results were used as in the study.

3.6.2 Data pre-processing steps

The study checked for missing values, duplication, and other potential errors. Missing values reduces the sample size, bias the results, reduce the precision and efficiency of statistical analysis, and lead to the loss of valuable information. In cases where the missing value was on the dependent variable (diagnosis of cervical cancer), the whole row was eliminated from the study.

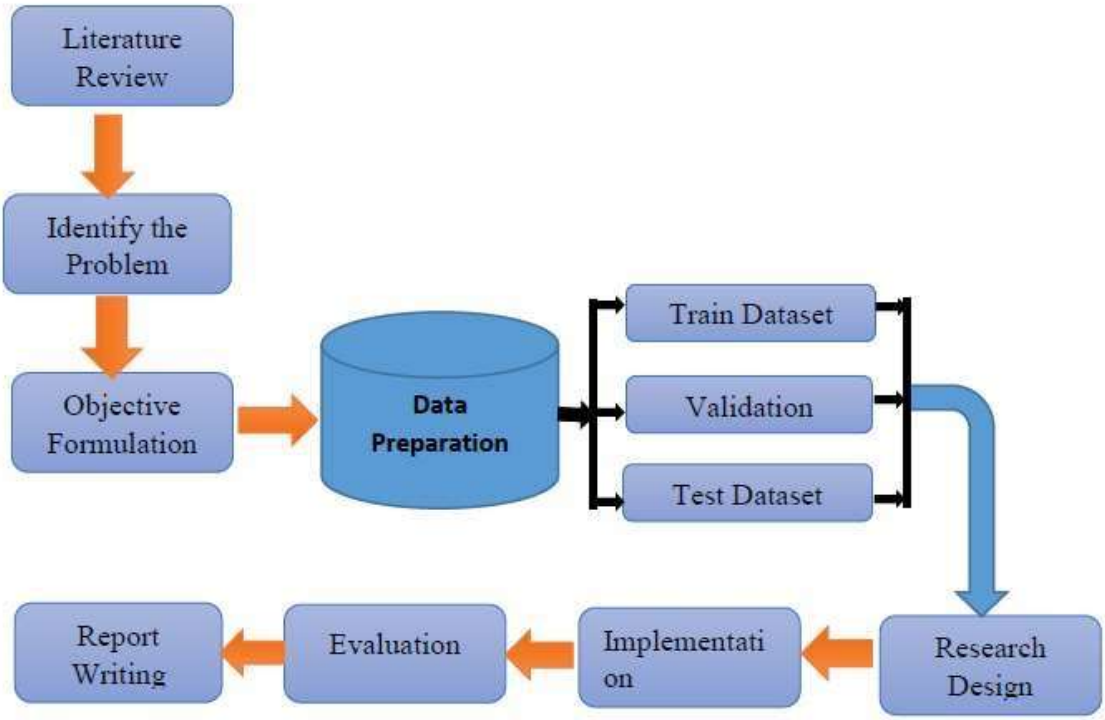

Figure 2: Research flow (Source: Author, 2024)

3.7 Model Development Flow

After the process of data management, the data was divided into training set and testing set. The development process started with preparing the data that was to be used for training the 5 different machine learning model and testing their performance on the test

set. The collected dataset was pre-processed with the target variable 'cervical cancer' defined. The data was split and the training set used to train the model. The feature engineering and other techniques was used during the training stage and the model trained be evaluated with the test set and their inference exported.

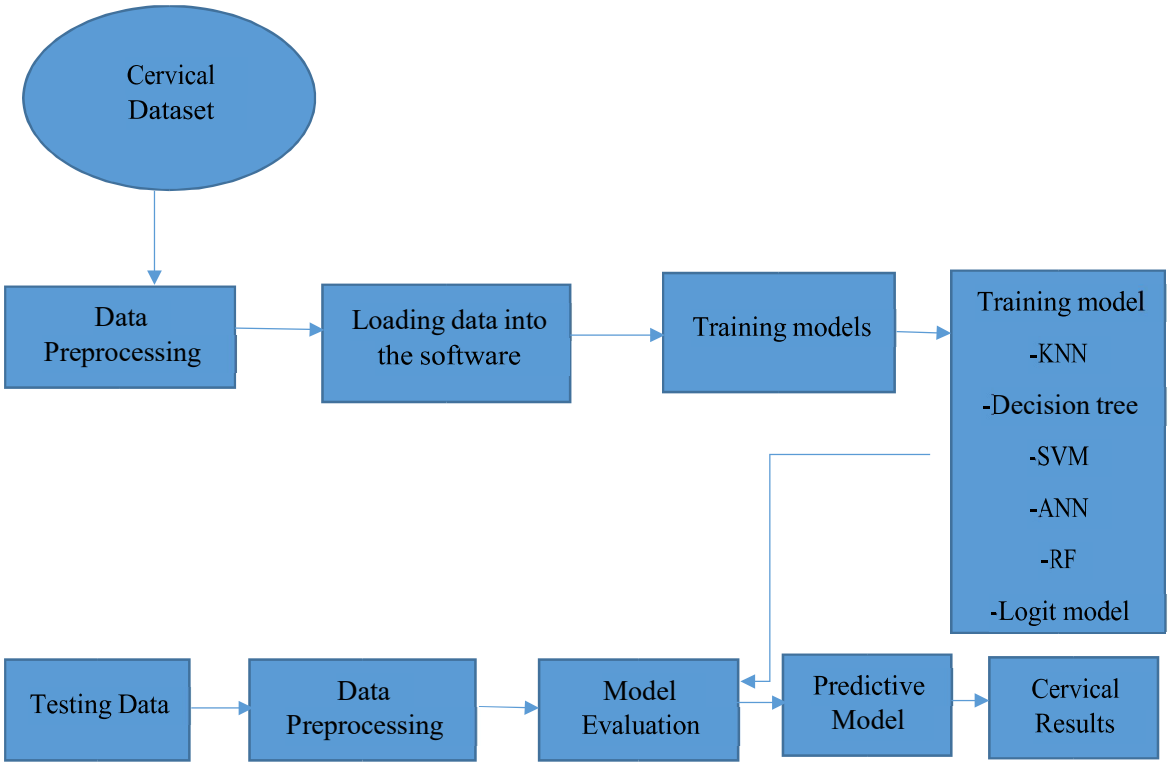

Figure 3: Detection of cervical cancer model development stages

(Source: Author, 2024)

3.8 Handling Class imbalance

3.8.1 The SMOTE Method

The synthetic minority oversampling technique (SMOTE) was used to balance the training data for the artificial neural network model. According to Chawla et al. (2002), SMOTE generates synthetic samples for the minority class group which increases the representation of the minority class by interpolating between the minority classes that exists. The SMOTE method was picked because of its complexities and sensitivity when dealing with class imbalance which could have led to poor generalization of the classes that are underrepresented (Fernández et al., 2018). SMOTE improves the artificial neural network's ability to learn patterns that are meaningful from minority classes which enhances the ANN's predictive performance for cervical cancer prediction by creating a balanced training dataset. Japkowicz and Stephen (2002) indicated that SMOTE mitigates the risk of overfitting to the majority class which is an issue that is common in artificial neural network trained on the imbalanced medical dataset. SMOTE is partially effective when dealing with datasets with high-dimension such as the cervical cancer dataset where the synthetic samples are helpful in capturing the distribution of the minority classes without the introduction of other excessive noise (Batista et al., 2004).

3.8.2 The Class Weighting

Class weighting method with the balanced parameter was used for the Random forest, logistic regression, logistic regression, decision tree and support vector machine to address the imbalance in the cervical cancer dataset where the positive cervical cancer cases were underrepresented (minority class). The class weighting method assigns weights to the minority class (positive cervical cancer cases) during the model training period inversely

proportional to the class frequency which ensures that the model is prioritizing learning from the minority class without interfering with the structure of the dataset. The approach of using class weighting preserves the original distribution of the dataset and it is also computationally efficient which makes it suitable for linear and tree based models where the imbalances in the class could have biased the prediction towards the majority class (He & Garcia, 2009). The class weighting enhances the models ability to detect critical cases that are rare such as cervical cancer by adjusting the loss function to penalize the misclassifications of the classes that are fewer more heavily (Burez & Van den Poel, 2009). This method was well suited for the cervical cancer dataset with huge imbalance as it avoids the risk of overfitting that is associated with data manipulation technique (Chawla et al., 2002).

3.8.3 Data Partitioning for ML predictions

The dataset was divided into two parts: training, and testing sets. The training set was part of the data used to train the detection and classification model, while the testing set was used to test the performance of the models after they had been trained. The training set was used to evaluate the model performance and pick the best performing model. The dataset was partitioned into training and testing sets following the 70-30 rule which is a widely accepted guideline (Smith et al., 2018). The training set occupied 70% of the total dataset. The testing set occupied 30% of the original set.

3.8.4 Feature Selection and Extraction

This study used the existing literature review to select relevant variables used in the classification model. The study reviewed some of the risk factors used in the previous study

in the same area to provide insights into significant variables. Understanding this risk factor helped to guide in choosing the factors associated with cervical cancer.

Seeking Expert Knowledge: This study sought input from the domain experts to ensure that the questionnaire captured clinically relevant risk factors for cervical cancer comprehensively. Domain knowledge is understanding and knowledge in a specific discipline or field (Feltovich et al., 1997). The oncologist, epidemiologist, and public health professionals are some of the domain experts that were consulted. These experts provided insights into the risk factors associated with cervical cancer based on their experience. A preliminary questionnaire draft was reviewed by the experts in iterative rounds where revisions were made to align it with cultural applicability and clinical relevance to systematically incorporate the expert input. The process of seeking domain knowledge ensured that the questionnaire was valid and it adhered to real world epidemiology patterns.

This study used the quality and availability of the data to select the features that were to be included in the model. The information recorded in the healthcare facilities and what the individuals are comfortable giving were used.

The study employed statistical techniques such as correlation and chi-square tests to determine the most significant features that affect cervical cancer. This study used feature selection algorithms: Lasso regression (Lockhart et al., 2014) and information gain methods (Kieffer, 2006) to determine the most important predictor variables in the prognostic model. These algorithms eliminate features depending on their importance in the detection model.

The study considered domain-specific considerations specific to Western Kenya. Lifestyle patterns, genetic predisposition, and environmental factors gave insight into the features that could be included in the predictive model.

This study used principal component analysis to reduce the dimension in the data before fitting a prognostic model. These techniques were used to transform the original features of the data that were collected into a lower-dimension representation while maintaining the most critical information. The principal component analysis is a dimension reduction method that is often used to reduce the dimensionality of large data sets where most of the variation in the data can be described by fewer dimensions (Dunteman, 1989).

3.9 The Model Development

This study aimed to develop a reliable and accurate detection model for predicting cervical cancer cases in Western Kenya that could be used for screening and prevention efforts. The pair of factors and response variable from the training set were used to train the various machine learning model. The model learned the underlying patterns and relationships in the train data by iteratively adjusting its parameters. The development process involves several machine learning model with each offering a different perspective on the likelihood of having cervical cancer.

3.9.1 Logistic Regression model

A logistic regression model which is a linear model that was used to predict the probability of cervical cancer based on the linear combination of the input features from the cervical cancer dataset. A logistic regression model was chosen for this study since it is easier to

interpret which makes it more suitable for making decisions clinically and it is able to handle binary classification problems like cervical cancer prediction.

Data processing

The numeric features were standardized to have a mean of 0 and a standard deviation of 1 which ensured that each variable contributed equally to the model. The categorical features were one-hot encoded to binary variables which made it easier to interpret their results and fit the model.

The class imbalance were fitted using class weighting which assigned 5:1 weight to the minority class (positive cancer cases) so that the model could penalize the misclassification of cervical cancer cases heavily.

Feature selection

The lasso regression was used to select key features so as to reduce instances of multicollinearity and also to improve the model interpretability (Li et al, 2022). The lasso regression applied an L1 penalty shown by equation ii below to shrink the features that were less important (coefficients that were close to zero):

$$Minimize \frac{1}{2n}\sum_{i=1}^{n}\left(yi - \left(\beta_0 + \sum_{j=i}^{p}\beta_j x_{ij}\right)\right) + \lambda\sum_{j=i}^{p}|\beta_j| \qquad \text{ii}$$

Where λ = 0.01 was chosen via cross-validation so as to balance the feature selection and also for the model fit.

Model Architecture

The logistic regression model was implemented using the python's scikit-learn's LogisticRegression class with the 'lbfgs' solver. The 'lbfgs' Solver optimized the log-likelihood function. This was done using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm.

The logistic regression model predicted the probability of cervical cancer using the function:

$$P(y_i = 1|Xi) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \cdots + \beta_k * X_{i12})}} \quad \text{iii}$$

Where:

y_i= 1 if the patient has cervical cancer, y_i = otherwise

β_0 *is the intercept*

$\beta_1 \ldots \beta_{12}$ are the coefficient of the selected features

X1, X2, X3... X12 are the predictors

Training process

The logistic regression model was trained on the 70% training set using the maximum likelihood estimation (MLE). This was to maximize the log-likelihood:

$$Log - likelihood = \sum_{i=1}^{n} [y_i log(p(y_i = 1|X_i)) + (1 - y_i) log(1 - p(yi = 1|X_i))] \quad \text{iv}$$

The class weight was set to balance by adjusting the weights inversely proportional to the class frequency (10 (Positive cases):1 (Negative cases)). The study used the 5-fold

validation to assess the generalization performance with the following regularization parameters:

C = 1.0 (the inverse of the regularization strength) which was selected from *c = [10, 1, 0.1, 0.01, 0.001]* via the grid search with the aim of minimizing the validation error.

Hyperparameter Justification

c = 1.0: This balanced the model complexity and regularized the model to prevent overfitting while maintaining its predictive power. The maximum iteration of 1000 ensured that there was convergence for the optimization algorithm while the random seed of 42 was used to ensure reproducibility of the results.

Final Model

The mathematical equation of the final model was:

$$\begin{aligned} Logit\,(p) = {} & \beta_1 + \beta_2 * Age + \beta_3 * Sexual\ partners + \beta_4 \\ & * First\ sexual\ intercourse + \beta_5 \\ & * Number\ of\ pregnancies + \beta_6 * Smokes\ (Year) \\ & + \beta_7 * Hormonal\ contraceptives\ (Year) + \beta_8 \\ & * Hormonal\ contraceptives + \beta_9 * PID + \beta_{10} * HIV \\ & + \beta_{11} * Pap\ smear + \beta_{12} * IUD \end{aligned} \quad \text{iv}$$

3.9.2 The Decision Tree model

Data processing

There was no scaling since decision tree is an invariant to feature scales so the numeric features in the data were not standardized, reducing the computational costs and the preprocessing complexities.

The class imbalance was handled using class weighting method during model training. The class weights were set to $W_1 = 10$ for the positive cervical cancer cases and $W_0 = 1$ for negative cervical cancer cases. The class weighting method was also chosen in this study since it maintains the original data distribution. This penalized the misclassifications of the positive cancer cases more heavily balancing the dataset. This study chose the 10:1 to approximate the inverse class ratio and adjusted downwards to balance the specificity and sensitivity while at the same time avoiding the model bias towards positive predictions.

Feature selection

There was no explicit feature selection technique that was used when fitting a decision tree model. This is because decision tree inherently perform feature selection through the splitting criteria. All the 15 features in the data were included when fitting the model so as to enable the algorithm to prioritize features that reduces impurity in a more effective manner. The feature importance was calculated post training as the total weighted Gini Impurity reduction across the splits where each feature were involved:

$$Importance(X_j) = \sum_{s\epsilon\ splits\ on\ x_j} \Delta Gini_{weighted}(D,s).\frac{\sum_{k\ \epsilon\ D} w_{yk}}{\sum_{k=1}^{n} w_{yk}} \quad \text{v}$$

Where $\Delta Gini_{weighted}$ is the impurity reduction for split (s)

(D): is the node,

w_{yk}:is class weights

This approach highlighted the features in the data that were clinically relevant without the preprocessing bias.

Model Architecture

The decision tree is a hierarchical classifier which partitions the feature space recursively into regions based on the feature threshold. This culminated into leaf nodes that assigned a class as either negative (0) or positive (1). The decision tree was constrained to:

Maximum depth of 3 to prevent overfitting and maximum sample per split as 5 to ensure we get a robust split.

Node splitting

Each of the node split was based on the weighted Gini Impurity that was adjusted for class imbalance:

$$Gini_{weighted}(D) = 1 - \sum_{i=0}^{1}\left(\left(\frac{w_i \sum_{k\epsilon D} I_{\cdot}(y_k = i)}{\sum_{k\epsilon D} w_{yk}}\right)\right)^2 \quad \text{vi}$$

Where $W_0 = 1$ *and* $W_1 = 10$

I is the indicator function

For the split (s) into left (D_L) and right (D_R) child nodes, the function for impurity reduction used was:

$$\Delta Gini_{weighted}(D, s) \quad \text{vii}$$

$$= Gini_{weighted}(D) - \frac{\sum_{k \epsilon Dl} w_{yk}}{\sum_{k \epsilon D} W_{yk}} Gini_{weignted}(D_l)$$

$$- \frac{\sum_{k \epsilon DR} w_{yk}}{\sum_{k \epsilon D} W_{yk}} Gini_{weignted}(D_R)$$

The split that maximizes $\Delta Gini_{weighte}$ was then selected.

Leaf node

The leaf node assigned the class that maximized the weighted sum using the function:

$$\overset{\wedge}{\gamma}.(D_{leaf}) = arg_{i \epsilon (0,1)}^{max}. \sum_{k \epsilon D_{leaf}} w_{yk} I(y_k = i) \quad \text{viii}$$

The tree could go up to the size of $2^5 = 32$ nodes given that it had a max depth of 5 and a stopping criteria with a minimum sample split of 5 which ensured that the tree had fewer nodes ensuring that the structure of the final model was compact and interpretable. The constraints (min_samples_split=5 and depth=5) balanced the tree model complexity and generalizations which was critical for the training dataset which was not extremely big.

Training process

The aim was to construct a decision tree that minimized the weighted classification error using the cost-adjusted Gini impurity by optimizing the splits.

The algorithm started with the root nodes which contained the 678 training samples

Recursive splitting

For each of the nodes that had at least 5 samples and depth that was less than 5:

i. All the possible splits on the 15 features were evaluated

ii. $Gini_{weighted}$ (D_L), $Gini_{weighted}$ (D_R), and $Gini_{weighted}$ (D, s) for each split were computed

iii. The split maximizing $\Delta Gini_{weighted}(D,s)\ was\ selected$

If there was no split that reduced the impurity, then it indicates that the node had fewer than 5 samples or depth of 5 and it was marked as a leaf.

Leaf assignment

The class were assigned to each leaf based on the weighted majority using the function:

$$\hat{\gamma}.\left(D_{leaf}\right) = arg_{i\epsilon(0,1)}^{max}. \sum_{k\epsilon D_{leaf}} w_{yk} I(y_k = i) \qquad \text{ix}$$

Hyperparameter tuning

5-fold cross validation to tune hyperparameters was performed where: max_depth[3, 5, 10] and min_sample_split[2, 5, 10]. 5-fold validation divided the training set into 5 folds (approximately 135 to 136 samples each) which ensured there was a robust hyperparameter selection.

Cost sensitive learning

The class weights $W_1 = 10$ for the positive cervical cancer cases and $W_0 = 1$ for negative cervical cancer cases adjusted the Gini impurity to prioritize the positive cancer cases ensuring that the split reduced the minority class:

$$Gini_{weighted}(D) = 1 - \sum_{i=0}^{1} \left(\left(\frac{w_i \sum_{k \epsilon D} I_{\cdot}(y_k = i)}{\sum_{k \epsilon D} w_{yk}} \right) \right)^2$$

Convergence

The decision tree was fully grown when all the nodes met the stopping criteria which was a depth of 5 and samples less than 5 or no impurity reduction. This resulted in a deterministic structure given the random seed that had for reproducibility of the model.

3.9.3 Random forest

The study also used a random forest model. Random forest is an ensemble method that combines many decision tree to improve the accuracy of prediction and robustness of the model (Breiman, 2001). The random forest model was chosen to predict cervical cancer due to its ability to handle high-dimensional noisy data and it also captures complex interactions.

Development process

Data processing

There was no scaling since the tree based model is invariant to feature scales. The class imbalance was addressed using class weighting with a 10:1 weight for the minority class (cervical cancer cases).

Feature selection

All the 15 features in the dataset were used to train the random forest model as the model can handle high-dimensional data and it also automatically ranked the feature importance

using Gini impurity. The feature importance were calculated after post-training to identify the predictors that are key.

$$Importance(X_j) = \frac{1}{T}\sum_{t=1}^{T} \sum_{s \in splits\ on\ X_{j\ in\ tree\ t}} \Delta Gini_{weighted}(D,s).\frac{\sum_{k \in D} w_{yk}}{\sum_{k=1} w_{yk}} \quad \text{xi}$$

Where:

(T): Number of trees

$\Delta \text{Gini}_{\text{weighted}}$: Is the impurity reduction for split (S)

(D): Is the node

w_{yk}: Class weight

n: Training sample size

Model Architecture

Figure 4 shows the model Architecture

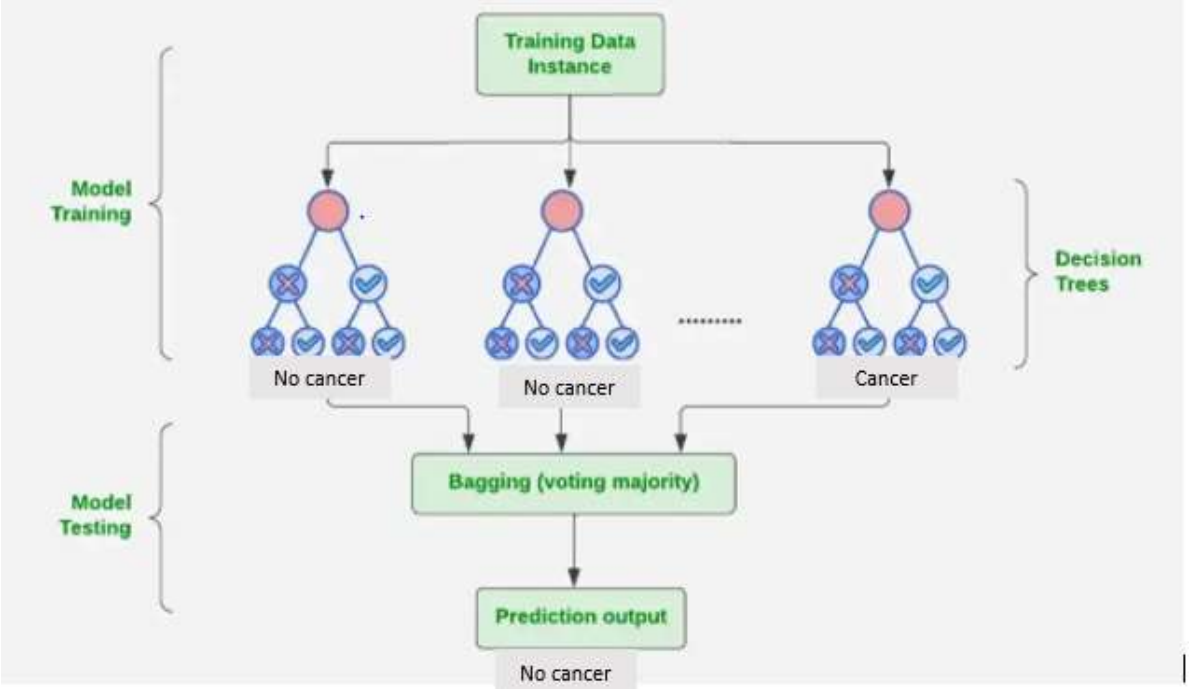

Figure 4: Random forest model Architecture ***(Source: Author)***

The random forest used was an ensemble of T =100 decision trees. Each of the decision tree was trained on a bootstrap sample of the training data and using a random subset of features. The final prediction was the majority of the vote across all the trees:

$$\hat{y}(x) = mode\{T_1(x), T_2(x), \dots T_{100}(x)\} \qquad \text{xii}$$

Where:

$T_1(x) \in \{0, 1\}$ Is the prediction of the (t)-th tree for input (X).

Tree construction

Each of the tree was grown with the following properties:

Maximum depth of 10 which limits the complexity and prevents overfitting.

Minimum sample per split of 5 to ensure there are robust splits.

4 Random feature subset at each node (sqrt {15) = approx. 4) to introduce diversity and reduce correlation.

Node splitting

Each of the node splitting was based on the weighted Gini Impurity after the adjustment of the class imbalances using the function:

$$Gini_{weighted}(D) = 1 - \sum_{i=0}^{1} \left(\frac{w_i \sum_{k \in D} I(y_k = i}{\sum_{k \in D}(w_{yk})} \right)^2 \quad \text{xiii}$$

Where $W_0 = 1$, $W_{10} = 10$

I is the indicator function

Leaf assignment

$$\hat{y}(D_{leaf}) = arg_{i \epsilon (0,1)}^{max} \sum_{k \epsilon \ leaf} w_{yk} I(y_k = i) \quad \text{xiv}$$

Training process

The objective of developing a random forest model was to minimize the weighted classification error approximated by the out-of-bay error using the cost adjusted Gini impurity.

The model created 100 decision trees. Bootstrap sampling (sampling with replacement) from the 678 training sample was conducted which resulted in approximately 63.2%. The remaining samples formed the OOB set for tree.

Hyperparameter tuning

5-fold cross validation to tune hyperparameters was performed where: max_depth[5, 10, 20], n_estimators [50, 100, 200] and min_sample_split[2, 5, 10]. 5-fold validation divided the training set into 5 folds (approximately 135 to 136 samples each) which ensured there was a robust hyperparameter selection. Based on cross-validation performance, the values n_estimators=100, max_depth=10, and min_samples_split=5 were selected to fit random forest.

3.9.4 Support Vector Machine

A Support vector machines was also used to fit a predictive model. Support vector machines aim to determine an optimum hyperplane that separates two classes with a maximum margin, making it perfect for this study as the study intend to separate the individuals into those with and without cervical cancer.

Feature selection

The principal component analysis was used to transform the 15 features in the dataset into 10 principal components:

$$X_{PCA} = X.W \qquad \text{xv}$$

Where:

W: The matrix of eigenvectors that was corresponding to the top 10 eigenvalues.

Model Architecture

Figure showing SVM model Architecture

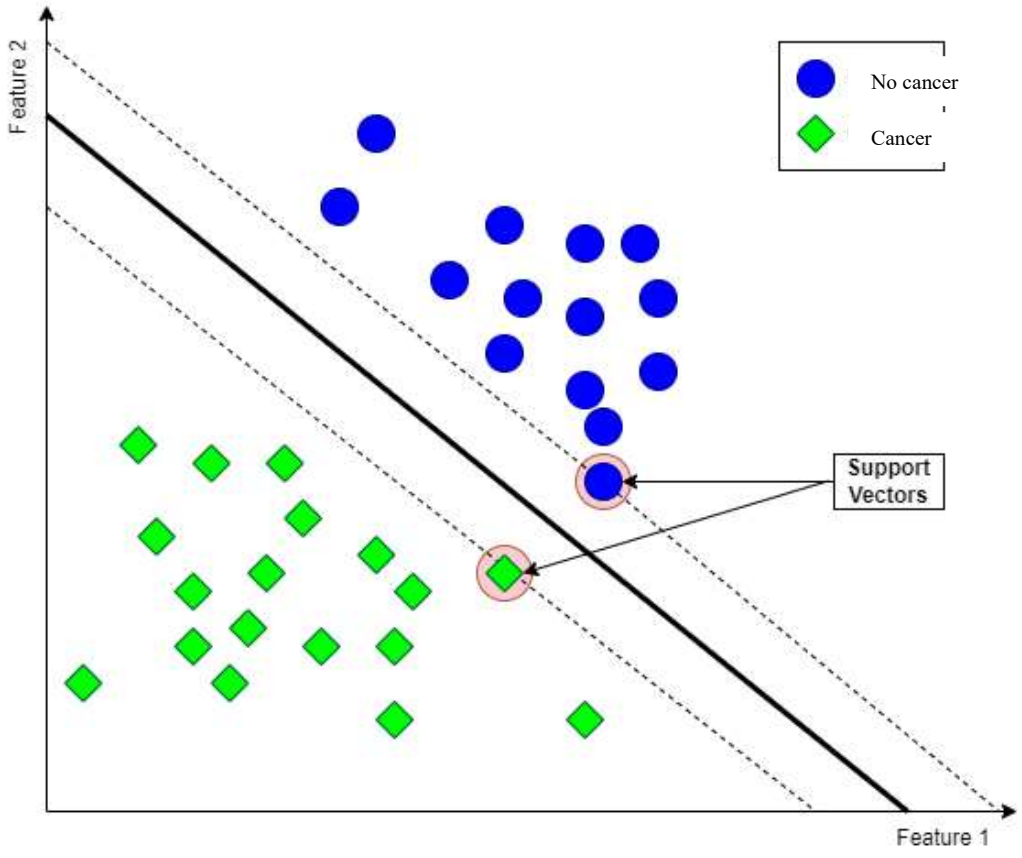

Figure 5: SVM model Architecture *(Source: Author)*

The Support vector machine determined a hyperplane that best separated the data points of two classes. The decision boundary for this binary classification was represented as:

$$W^t x + b = 0$$

Where:

- w = represents the weight vector (normal to the hyperplane, determining its orientation).

- x = represents the input feature vector such as age, HPV status, Sexual partners and Contraceptive for each individual among others.
- b = represents the bias term. The bias term determines the offset of the hyperplane from the origin.
- $W^t x + b$= represents the decision function where the sign indicates the class as being either positive or negative.

Optimization objective

The support vector machine maximizes the margin. The optimization problem is represented as:

$$Minimize \frac{1}{2}||w||^2$$

Subject to:

$$y_i(W^t x_i + b) \geq 1 \; for \; all \; i$$

Where:

Y_i = represents the class label where +1 for cancer and −1 for no cancer.

X_i = represents the feature vector for the (i)-th data point.

Model training

The support vector machine was trained on the training set using the RBF kernel. The hyperparameters were tuned using the grid search over: *C: [0.1, 1, 10]* and gamma: [0.001, 0.01, 0.1]. The best model used C=1 and gamma=0.01. This was because C=1 balances

margin maximization and classification error which was suitable for the noisy medical data like the one used in this study. The gamma=0.01 controlled the influence of individuals data points in the model which prevented overfitting in the high dimensional spaces. The RBF kernel captured the non-linear relationship in the cervical cancer dataset.

3.9.5 Artificial Neural Network (ANN)

An ANN was used in this project. This machine learning method was appropriate for this study since it had the ability to learn non-linear and complex relationships that exist in the dataset. With the advancement in computing ability of the R and Python software that were used in this study, it was easier to train and test the ANN.

Model Architecture

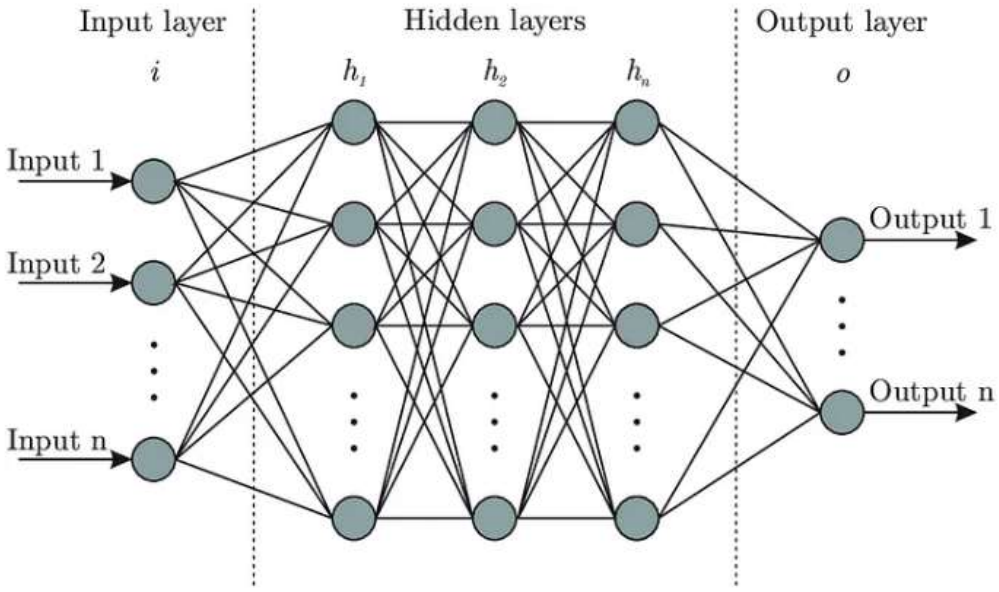

Figure 6: ANN model Architecture *(Source: Jain, Mao, & Mohiuddin, 1996)*

The artificial neural network was implemented using TensorFlow using three layers:

i. The input layer: 15 nodes which corresponded to the 15 features

ii. The hidden layer: 64 nodes with ReLU activation (f(x) =max (0, x))

iii. The output layer: Had 1 node with sigmoid activation $\sigma(x) = \frac{1}{1+e^{-x}}$ for classification of binary outcome.

The dropout rate of 0.2 was applied to prevent overfitting. The dropout rate of 0.2 Prevented overfitting by randomly dropping 20% of neurons during training.

The forward pass was computed using the function:

$$h = ReLu(W_1 x + b_1), \hat{y} = \sigma(W_2 x + b_2) \quad \text{xviii}$$

Where:

W_1 and W_2 are weight matrices

b_1 and b_2 are biases

Training process

The artificial neural network was trained using the Adam optimizer (with a learning rate of 0.001) with binary cross-entropy loss of:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y_i} + (1 - y_i)\log(1 - \hat{y_i})] \quad \text{xix}$$

The training of the model ran for 100 epochs with a batch size of 32 using the early stopping (patience=10) that was based on the validation loss. Batch size of 32 was suitable for stable gradient updates and also GPU memory constraints. The learning rate of 0.001 ensured that there was a stable convergence with the Adam optimizer.

The 5-fold validation was used in model training to assess the generalization of the model.

3.10 Evaluation Metrics

Once the different machine learning models had been trained on the training set, they were evaluated on the testing set. The testing set which constitutes 30% of the entire dataset acted as a benchmark for assessment of the model performance on unseen data. This study gauged the efficacy of the machine learning models fitted by exposing the models to this independent dataset. This dataset posed a new challenge to the model as the model encountered instances that it had not encountered during the training phase.

The evaluation part in the model testing component ensured that the model trained was not overfitting the training data and performing poorly on the test data. This ensures that the model is making predictions that can be relied upon.

Cervical cancer prediction was a classification problem. Confusion matrix is one of the most popular methods used to measure the performance of the classification model. Confusion matrix can be applied to both binary cases and multinomial cases. The confusion matrix shows the accuracy, 95% CI of the accuracy, kappa value, sensitivity, specificity, balanced accuracy and prevalence among other metrics.

In this study, accuracy represents the proportion of the cervical cancer cases that were correctly classified.

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{True\ Positive\ (TP) + False\ Positive\ (FP) + TrueNegative\ (TN) + False\ Negative\ (FN)} \quad \text{xx}$$

Precision represent the proportion of cervical cancer cases that were correctly predicted to be positive out of all cervical cancer predicted as positive.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad \text{xxi}$$

Sensitivity represent the proportion of cervical cancer cases that were correctly classified to be positive out of all the actual positive cervical cancer cases.

$$Sensitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad \text{xxii}$$

The F1 score represents the harmonic mean of precision and recall providing a more balanced measure for recall and precision. The specificity represent the proportion of cervical cancer cases that were negative that were correctly predicted as negative.

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + False\ Positive\ (FP)} \quad \text{Xxiii}$$

Accuracy of the models can be misleading in cases where the classes were imbalanced. To address this, the study employed specificity, sensitivity, and precision metrics to provide more details about the performance of the models. Given that the proportion of people with cervical cancer and those without is not same, there was an imbalanced which necessitate the use of the other metrics from the confusion matrix for evaluation.

3.11 Ethical Considerations

i. The names and IDs of the respondents were not used to ensure anonymity and the confidentiality of the respondent's information.

ii. Each respondent was given written consent indicating their willingness to participate and decide to opt-out if they feel so without any restriction.

iii. Participation was voluntary.

iv. The study obtained permission and approval from the hospital administration and other bodies such as NACOSTI and the Ethical approval committee (MMUST ETHICAL COMMITEE).

CHAPTER FOUR

RESULTS

4.1 Introduction

This chapter presents results of each of the specific objectives as outlined below. The summary of the results include the descriptive statistics, cervical cancer screening and diagnosis reproductive and sexual health characteristics, cytology and biopsy results, screening and diagnosis results. On inferential statistics, the most suitable ML models were employed and fitness assessment was done using model performance metrics and evaluation.

4.2 Summary of Socio- Demographic, Clinical information and Behavioral Characteristics of Study participants

4.2.1 Social demographic and behavioral characteristics

Based on the results from Table 1, a majority of the women who participated in this study were mostly aged between 20 and 25 years (39.77%, n= 385) followed by those aged between 26 and 35 years (36.16%, n=350). A smaller portion of the participant were aged between 36 and 45 (12.40%, n=120), below 20 years (9.92%, n=96), 46-60 years (1.34%, n=13), and above 60 years (0.41%, n=4).

The majority of the participants reported that they were using hormonal contraceptives (55.68%, n=539) while only 31.71% (n=307) of the respondents did not use hormonal contraceptives. Only 12.60% of the participants did not indicate their contraceptive status.

In terms of smoking behavior, the data showed that 83.88% (n=812) of the participants were non-smokers while only 14.46% (n=140) were smokers.

Table 1: Socio- Demographic, Clinical information and Behavioral Characteristics of Study participants.

Characteristic	Frequency	%
Age Category		
Below 20	96	9.92%
20-25	385	39.77%
26-35	350	36.16%
36-45	120	12.40%
46-60	13	1.34%
Above 60	4	0.41%
Hormonal Contraceptives		
Yes	539	55.68%
No	307	31.71%
Not indicated	122	12.60%
Smokes		
No	812	83.88%
Yes	140	14.46%

4.2.2 Reproductive and Sexual Health Characteristics

This section summarized the reproductive and sexual health characteristics of the respondents who were involved in the study. These characteristics included the number of pregnancies, age at first sexual intercourse, and number of sexual partners.

Based on the results from Table 2, women on average in the study had 2.08 pregnancies. The standard deviation of 1.51 indicated that the number of pregnancies varied moderately around the mean. Women involved in the study on average had between 1 and 4 pregnancies (Mean±1 SD). The results also shows that on average women in the study first had sexual intercourse was 19.79 years. The SD of 3.26 years indicated a moderate spread in the age of first sexual intercourse. The data showed that women in Western Kenya likely had their first sexual experience between 16.53 and 23.05 years (mean ± 1 SD).

Table 2: Summary statistics table

Characteristic	Mean	SD
Number of pregnancies	2.08	1.51
First sexual intercourse	19.79	3.26
Number of sexual partners	2.48	1.70

4.2.3 Cervical Cancer Screening and Diagnosis

This section presents the findings related to cervical cancer screening and diagnosis, including Pap smear and biopsy results using data visualization methods.

Figure showing Cytology results distribution

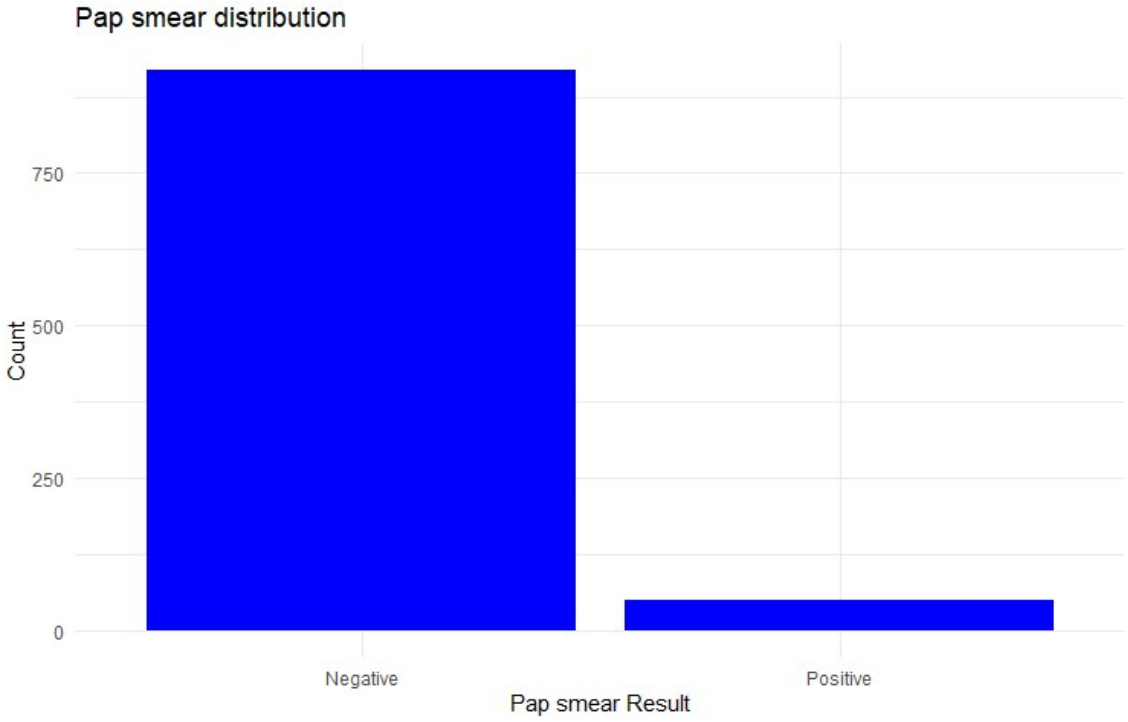

Figure 7: Distribution of Pap smear results

The results from figure 7 above showed that 94.83% (n=918) of the women had no abnormal Pap smear (Cytology) results, while 5.16% (n=50) had abnormal results.

Figure below shows biopsy results distribution

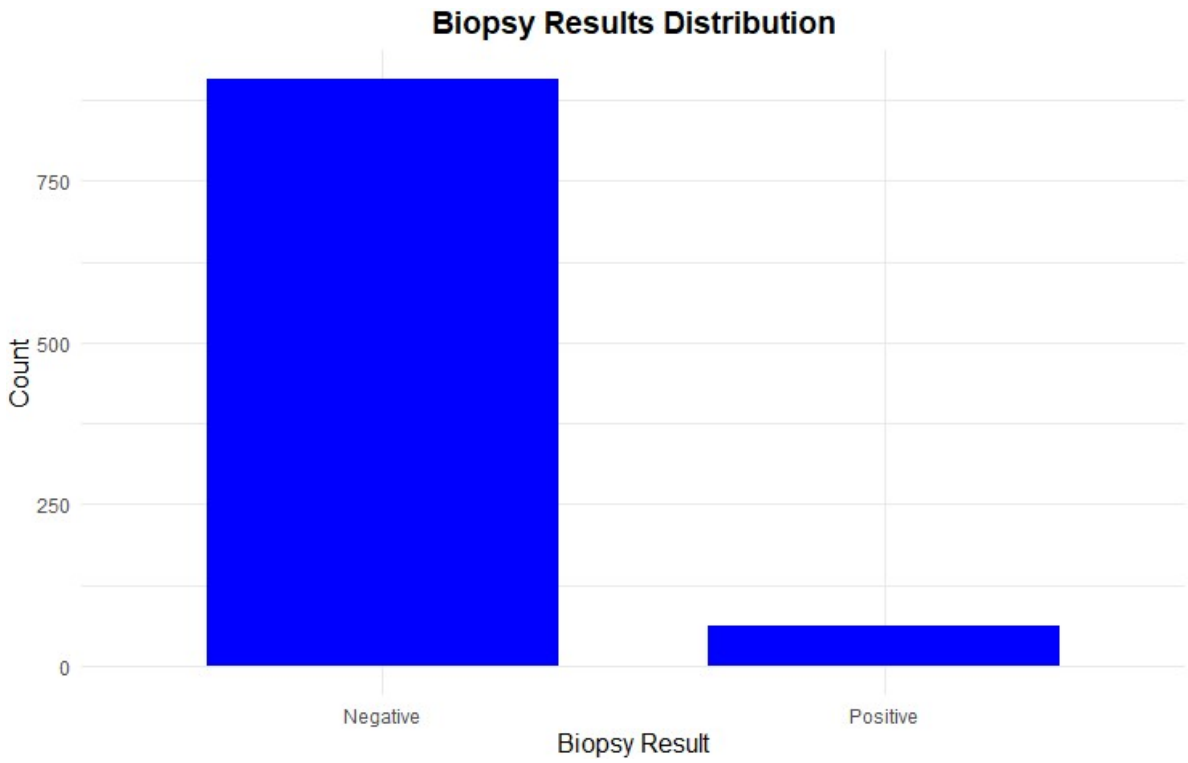

Figure 8: Distribution of Biopsy results

The results from figure 8 above showed that 93.70% (n=907) of the women had no biopsy-confirmed abnormalities, whereas 6.30% (n=61) had biopsy-confirmed abnormalities.

Bivariate relationship

Number of sexual partners and Biopsy results

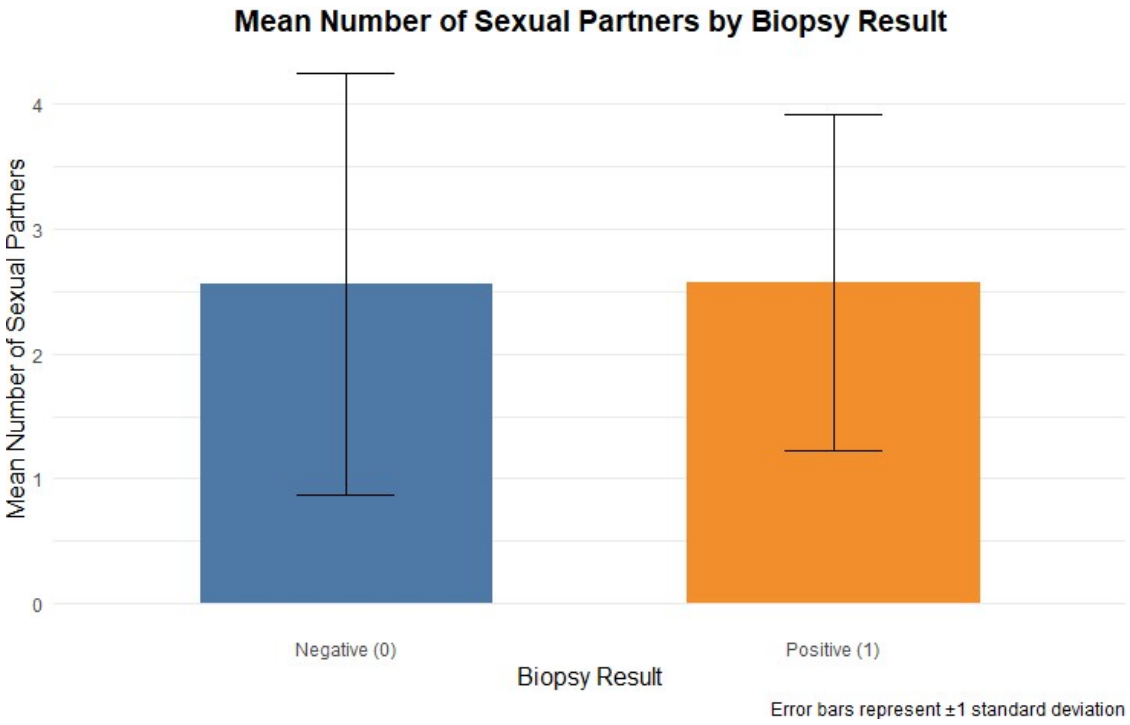

Figure 9: Mean number of sexual partners by cancer occurrence

The results from figure 9 shows that the average number of sexual partners a respondent had been involved with was higher among those with cervical cancer (M=2.57, SD=1.35) compared to those without cervical cancer (M=2.45, SD=1.68). The Mean difference does not show significance difference, this was determined by the logit regression model result (Table 3).

First sexual intercourse and Biopsy results

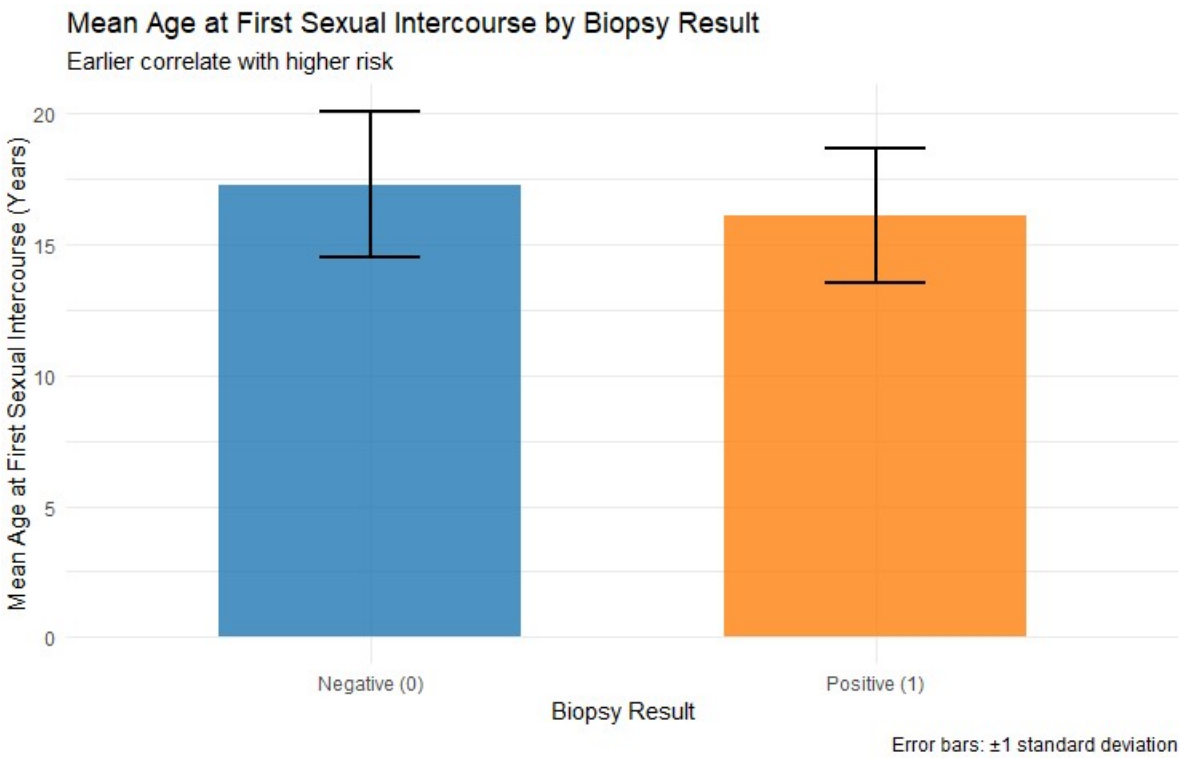

Figure 10: Mean first sexual intercourse by cancer occurrence

The results from figure 10 shows that participants not diagnosed with cervical cancer had a higher average age at first sexual intercourse (M = 17.3, SD = 2.80) compared to those with cervical cancer (M = 16.1, SD = 2.57). The findings from the logistic regression below showed that age at first sexual intercourse was marginally significant (p = 0.16)

4.3 A comparison of the predictive abilities of various developed machine learning models in detecting cervical cancer.

4.3.1 Model Development and Evaluation on unseen data

The performance of the machine learning models fitted- Logistic Regression, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) developed to predict cervical cancer were presented in this section. The models were trained on 70% of the dataset (n=678) and evaluated on 30% of the dataset (n=290) as shown in section 3.5 above. The evaluation metrics derived from the confusion matrix included accuracy, precision, recall and specificity. The results of each model was reported separately followed by a comparative analysis of the models. The ROC curve and confusion matrix below represent the performance of the model on unseen data (test data)

4.3.2 Model Performance

The performance of each model on the test set is detailed below, with key metrics summarized in Table 4.3. Each model's ability to classify cervical cancer cases ("Yes") versus non-cases ("No") is assessed, considering the dataset's imbalance (6.3% positive biopsy cases, per Section 4.3).

Logistic Regression

Table 3: Table showing the results of Logit regression model

	Estimate	Std. Error	p-value
Intercept	-4.60	1.13	0.00
Age	-0.02	0.03	0.04
Sexual partners	-0.06	0.13	0.65
First sexual intercourse	0.09	0.06	0.16
Num. Pregnancies	0.19	0.14	0.19
Smokes (years)	0.03	0.03	0.42
Hormonal. Contraceptive (Years)	0.08	0.04	0.04
Hormonal. Contraceptive	0.16	0.39	0.07
PID	-13.49	27.44	0.10
HIV	1.28	0.75	0.09
HPV	14.56	19.07	0.09
Pap-smear	2.62	0.39	0.00
IUD	0.79	0.67	0.23

Based on the results in Table 3 above, the fitted logistic regression model equation is as shown below:

Model equation

Logit (p) =−4.60−0.02**Age*−0.06* *Sexual partners* +0.09**First sexual intercourse* + 0.19* *Num. Pregnancies*+0.03**Smokes (Years)* + 0.08* *Hormonal. Contraceptive (Years)* + 0.16* *Hormonal Contraceptive (Yes)* − 13.49* PID + 1.28**HIV* + 14.56* *HPV* + 2.62* *Pap-smear* + 0.79* *IUD*

Logit model performance metric

Table 4: Logit model performance metrics

Metric	Value
Accuracy	0.8247
Precision	0.1842
Specificity	0.7000
Sensitivity	0.8315

Based on the results from Table 4 above, the logit model achieved an accuracy of 82.47% on the unseen data. This indicates that out of all the test samples, 82.47% were correctly classified as either having cervical cancer or not having cervical cancer. The precision value was 0.1842. This indicates that the model correctly predicted positive cervical cancer cases 18.42% of the time. The model additionally recorded a sensitivity of 70.00%. This meant that the logit model correctly identified 70.00% of the actual positive cervical cancer cases. The specificity was 83.15%. This indicated that 83.15% of the actual negative cases (no cervical cancer) were correctly classified by the logistic model.

Decision tree

Tree plot

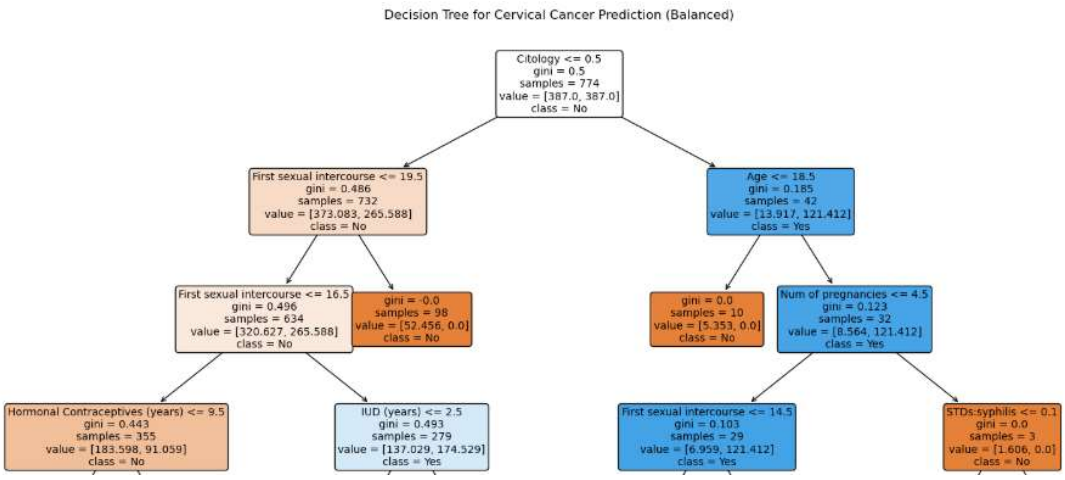

Figure 11: Decision tree plot (Source: Author, 2024)

The decision tree for cervical cancer showed how the different factors contributed to the classification of cancer. The root node split on Pap smear (<=0.5) which means that Pap smear results significantly influenced cervical cancer prediction. If the pap smear was negative, the model further split based on first sexual intercourse age with earlier ages at a higher risk of having cancer,

Performance metric of Decision tree

Table 5: Decision tree performance metrics

Metric	Value
Accuracy	92.78%
Precision	33.33%
Specificity	95.65%
Sensitivity	40.00%

The decision tree model achieved an accuracy of 92.78 which indicates that the decision tree correctly classified test cases as either having or not having cervical cancer 92.78% of the time. The model had a sensitivity of 40% and a specificity of 95.65%. The precision of the model was 33.33%.

Random forest

Table 6: Random forest model performance metrics

Metric	Value
Accuracy	94.33%
Precision	40.00%
Specificity	98.37%
Sensitivity	20.00%

The random forest model achieved an accuracy of 94.33%, which indicates that it correctly classified test cases as either having or not having cervical cancer 94.33% of the time. The model had a sensitivity of 20.00% and a specificity of 98.37%. The precision of the model was 40.00%.

Support Vector Machine (SVM)

Table 7: SVM model performance metrics

Metric	Value
Accuracy	0.8299
Precision	0.1515
Specificity	0.8478
Sensitivity	0.5000

Based on the results from Table 7 above, the Support Vector Machine Classifier achieved an accuracy of 82.99% on the unseen data. This indicates that out of all the test samples, 82.99% of them were correctly classified as either having cervical cancer or not having cervical cancer. The precision value was 0.1515 which indicated that the model correctly predicted positive cervical cancer cases 15.15% of the time. The model additionally recorded a sensitivity of 0.50. This meant that the Support Vector Machine Classifier correctly identified 50.00% of the actual positive cervical cancer cases. The specificity of the Support Vector Machine Classifier was 0.8478. This indicated that 84.78% of the actual negative cases (no cervical cancer) were correctly classified by the Support Vector Machine Classifier. The training and evaluation of the Support Vector Machine Classifier took 1.5 seconds which reflected the moderate computational demand that is due to the hyperplane optimization.

Artificial Neural Network (ANN)

Table 8: ANN model performance metrics

Metric	Value
Accuracy	0.8660
Precision	0.1000
Specificity	0.9022
Sensitivity	0.2000

Based on the results from Table 8 above, the Artificial Neural Network (ANN) achieved an accuracy of 0.8660 on the unseen data. This indicates that out of all the test samples, 86.60% of them were correctly classified as either having cervical cancer or not having cervical cancer. The precision value was 0.10 which indicated that the model correctly predicted positive cervical cancer cases 10.0% of the time. The model additionally recorded a sensitivity of 0.20. This meant that the Support Vector Machine Classifier correctly identified 20% of the actual positive cervical cancer cases. The specificity of the Artificial Neural Network (ANN) was 0.9022. This indicated that 90.22% of the actual negative cases (no cervical cancer) were correctly classified by the Artificial Neural Network (ANN). The training and evaluation of the artificial neural network Classifier took over 3 seconds which reflected the complex model architecture.

4.3.3 Comparative Analysis

Comparative analysis results are summarized in the Table 9 below:

Table 9: Models Comparative analysis

Model	Accuracy	Precision	Sensitivity	Specificity
Logit	82.47%	18.42%	70.00%	83.15%
ANN	86.60%	10.00%	20.00%	90.22%
Random forest	94.33%	40.00%	20.00%	98.37%
SVM	82.99%	15.15%	50.00%	84.78%
Decision Tree	92.78%	33.33%	40.00%	95.65%

According to the results from Table 9 above, the performance of the machine learning models (Logistic Regression (Logit), Artificial Neural Network (ANN), Random Forest, and Support Vector Machine (SVM)) was evaluated on unseen test data to predict cervical cancer. The various performance metrics (accuracy, precision, sensitivity and specificity) revealed the distinct trade-off of the predictive models fitted in this study. Each of the model that was fitted in this study addressed the class imbalance (~6.3% positive cases). The class imbalance was addressed using class weight in logit, SVM and random forest model. The ANN employed SMOTE for oversampling of the minority class.

The random forest achieved the highest accuracy of the four models fitted with an accuracy of 94.33%. This indicated that 94.33% of all the test samples were correctly classified as either having cervical cancer or not having cervical cancer. The superior accuracy displayed by the random forest show how this model leveraged on its ensemble approach to handle the imbalance and complexity in the dataset. The random forest model vastly

classified a majority of the tests that were negative cases (98.37% specificity). The sensitivity of this model was however only 20.00% which means that it only identified 20% of the actual positive cervical cases. The precision of this model was 40% which indicated that when it predicted a positive cervical cancer case, it was correct 40% of the time. The random forest had the highest precision among the models.

The artificial neutral network recorded an accuracy of 86.60% which means it correctly classified 86.60% of the test samples. This model achieved a high specificity of 90.22% indicating the ability of ANN in identifying negative cervical cancer cases. The sensitivity of ANN was however notably low at 20% which indicated that ANN detected only 20% of the true positive cancer cases, similar to the random forest model. Precision for ANN was the lowest among the models at 10%. This means that only 10% of the positive cervical cancer predictions were accurate. This highlight the high rate of false positives. The poor sensitivity and precision may be as a result of overfitting to the SMOTE-balanced training data.

The logistic regression model (Logit) attained an accuracy of 82.47%. This indicated that 82.47% of the test samples were correctly classified as having or not having cervical cancer. The model demonstrated a sensitivity of 70% which is the highest among the four models fitted. This indicated that the model correctly identified 70% of the actual cervical cancer cases. The precision of the model was 18.42% which indicated that only 18.42% of the positively predicted cancer cases were correct. Specificity of the model was 83.15% which showed a reasonable performance in classifying negative cancer cases.

The support vector machine classifier achieved an accuracy of 82.99% which was slightly higher than the logit model. This indicated that 82.99% of the test cases were either

classified correctly as having or not having cervical cancer. The sensitivity was 50.00% which indicates that the model identified half of the true positive cancer cases which was a moderate performance compared to the logit's 70%. The precision was 15.15% which is lower than the logit model. This suggested that only 15.15% of the models positive prediction were accurate which again pointed to a high false positive rate. The specificity of the model was 84.78%. The specificity was 84.78% which was comparable to Logit. This indicated effective classification of negative cancer cases by this model.

The decision tree recorded an accuracy of 92.78% which means that it correctly classified 92.78% of the test samples as either having or not having cervical cancer. The model achieved a sensitivity of 40% which indicates that 40% of the true positive cervical cancer cases were identified correctly better than the Random forest and ANN model but below the SVM and Logit model. The precision was 33.33% which means that 33.33% of the model positive cervical cancer predictions were correct. The model outperformed a Logit, SVM, and ANN model but trailed a Random Forest model. The model had a specificity of 95.65% which indicated its strong performance in ruling out negative cancer cases but its performance was below that of a random forest model.

Comparatively, random forest excelled better in overall accuracy and specificity making it more reliable for ruling out cervical cancer. The low sensitivity limits the random forest from being used as a screening tool for detecting positive cervical cancer cases. The logit model outperformed the other models in sensitivity (70%). This made it more effective at identifying positive cervical cancer cases despite its lower precision (18.42%). The support vector machine had a balanced sensitivity (50%) and specificity (84.78%) but the second lowest precision. The Artificial neural network underperformed in sensitivity

(20%) and precision (10%). These results suggested the trade-off among the models. The logit and SVMmachine model prioritized sensitivity for screening while the ANN, Decision tree and Random forest model were better suited for confirmation rather than being used for initial detection of cervical cancer in imbalanced cervical cancer context.

4.3.4 Visualization of Model Performance

Receiver Operating Characteristic (ROC) curves for all the five models.

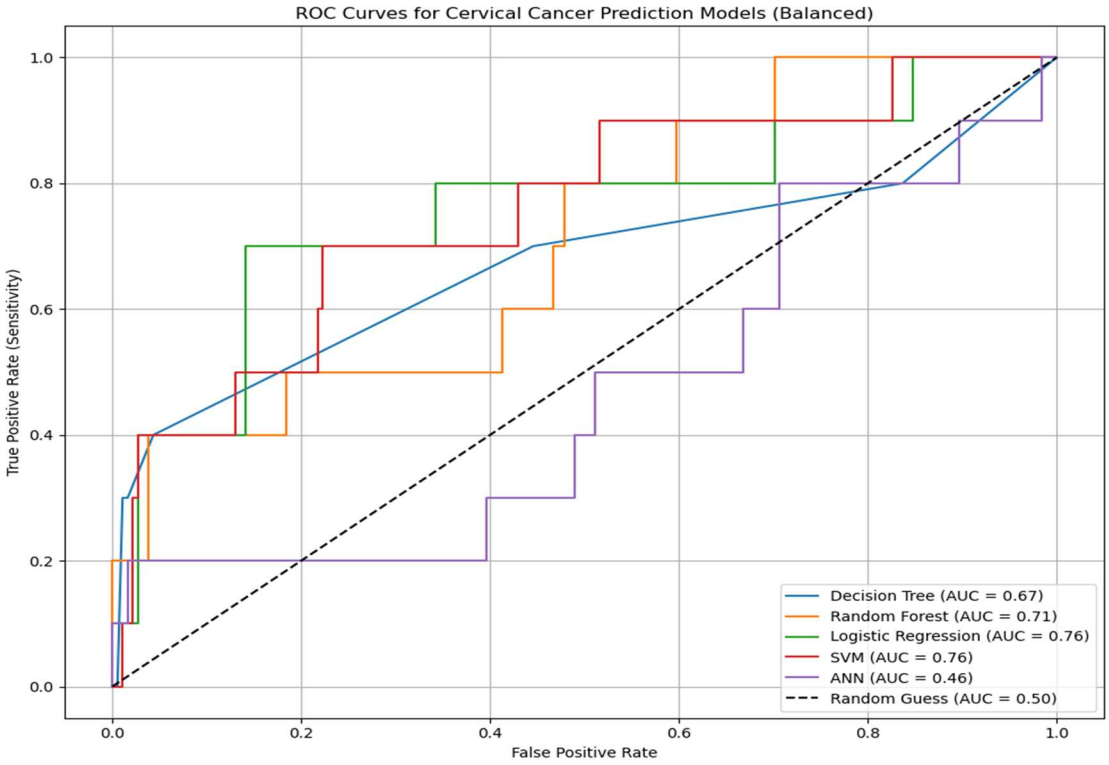

Figure 12: ROC plot

The Roc Curve was also used to compare the performance of the different models for cervical prediction. The logistic regression and the support vector machine (AUC=0.76) performed better than the other models which indicates that they are most reliable classification. The random forest (AUC=0.71) also performed well but slightly lower than

the logistic regression model and SVM. The decision tree (AUC=0.67) showed moderate performance while the ANN (AUC=0.46) performed the worst of the five models fitted. The Random Forest, SVM and logit model's superior AUC underscores their discriminative power.

4.4 Identifying the Key Risk Factors Associated with Cervical Cancer among Women in Western Kenya

This section highlights the significance of selected potential factors and compares the predictive capabilities of various models.

4.4.1 Feature importance and selection outcome

Table 10: Feature importance

Model		Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Decision tree	**Var**	Age	Pap smear	First sexual intercourse	IUD (Years)	Sexual partners
	GV	0.29	0.25	0.17	0.09	0.06
Random forest	**Var**	Pap smear	Hormonal contraceptive	Age	First sexual intercourse	Sexual partners
	GV	0.18	0.16	0.13	0.10	0.09
Logit model		Pap smear	HIV	IUD	Hormonal contraceptive	Sexual partners
SVM	**Var**	HIV	Pap smear	Hormonal contraceptive	STD (syphilis)	IUD
	GV	0.00054	0.00047	0.00032	0.00021	0.00021
ANN	**Var**	Hormonal contraceptive	IUD (Year)	Smoke (Years)	Pap smear	IUD
	GV	0.034	0.031	0.014	0.013	0.011

Using the Gini importance to measure impurity reduction, the decision tree identified Age (Gini Value = 0.29), Pap smear (Gini value = 0.25) and First sexual intercourse (Gini Value = 0.17) as the top predictors of cervical cancer. The number of sexual partners and IUD years also had a significant impact on cervical cancer. The importance of age and Pap smear aligns with the clinical expectations as age is known risk factor for cervical cancer and Pap smear results indicates cervical abnormality.

Also using Gini importance, Random forest ranked Pap smear (Gini value = 0.18), Hormonal contraceptive (Gini Value = 0.16), and Age (Gini Value = 0.13) as the most important features. The number of sexual partners and first sexual intercourse also followed closely. The ensemble nature of the random forest which averaged 100 trees reinforced the importance of Pap smear and Hormonal contraceptives was also with the clinical literature (section 2.2). Feature importance was assessed using coefficients and p-values for the Logit model (Table 3). The top variables included Pap smear, HIV, IUD, Hormonal contraceptive, and Sexual partners. The feature importance for SVM was derived via permutation importance due to the linear kernel and it identified Pap smear (Gini Value = 0.00047), and Hormonal contraceptives (Gini Value = 0.00032) as the top two predictors. Artificial neural network feature importance was assessed via permutation importance and it ranked Hormonal contraceptive (Gini Value = 0.034) and IUD (Years) as top two feature. Pap smear (Gini value = 0.031) was also important.

Cross model comparison

Across the five models that were fitted in this study, Pap smear emerged as the most consistent predictor. It ranked top three for Decision tree, Random forest, Support vector machine, and Logit and fourth in artificial neural network. The consistency of Pap smear

underscores its relevance clinically as a direct indicator of cervical cancer (Section 2.2). Hormonal contraceptive was also prominent as it ranked highly in Random forest, SVM, ANN and Logit model. This supports its role as a risk factor due to prolonged exposure.

Sexual partners appeared in the top five for Decision Tree (GV = 0.06), Random Forest (GV = 0.09), and Logit, reflecting its association with HPV transmission risk. HIV was notable in SVM (GV = 0.00054) and Logit (p = .09), aligning with its known impact on cervical cancer susceptibility. Features like Age (Decision Tree GV = 0.29, Random Forest GV = 0.13) and First sexual intercourse (Decision Tree GV = 0.17, Random Forest GV = 0.10) were more model-specific, indicating varying sensitivity to demographic factors.

CHAPTER FIVE

DISCUSSION

5.1 Introduction

A machine learning model was developed in this study to detect cervical cancer in Western Kenya to address the high burden of cancer in this low-income region in Kenya. The findings from Chapter 4 above offer insights into the participant's characteristics, key risk factors and the performance of machine learning models compared to traditional screening methods like Pap smear and HPV testing. This section discusses the results in the context of the hypothesis, existing research and their implications for improving cervical cancer screening in Western Kenya.

5.1.1 Participant Characteristics and Cervical Cancer Risk Factors

The majority of women who participated in the study were young, with 76% (735 out of 968) aged between 20 and 35 years, 40% aged between 20 and 25 and 36% aged between 26 and 35 (Table 1). These results matches the patterns in Sub-Saharan Africa, where cervical cancer affects younger women often due to early sexual debut and exposure to human papilloma virus (HPV), a major cause of cervical cancer (World Health Organization, 2022; Arbyn et al., 2020). Over half of the women involved in the study (56%) were using hormonal contraceptives which is a common practice in the region too, with only 14% classified as smokers which is a less common practice among African women compared to those from developed and wealthier nations (Bruni et al., 2018; Torre et al., 2015).

The study also found out that on average women had two pregnancies and their first sexual activity occurred around the age of 20 years (Table 4.2). The women with cervical cancer tended to have their first sexual experience at a slightly earlier period (around the age of 17 years) compared to those without cervical cancer (around the age of 18 years). The small difference in age at sexual debut between those with the disease and those without suggest that starting sexual activity at a much younger age increases the chances of HPV infection which is a known risk factor of cervical cancer (Plummer et al., 2016; LaVigne et al., 2017). Although this finding was not statistically strongly conclusive (p=0.16), it highlighted the need to focus on younger women to receive human papillomavirus vaccination and early screening in Western Kenya (Matenge & Mash, 2018).

The results shows that only 5% of the women had Pap smear and 6% had no biopsy (Table 1) which reflected the limited availability of the screening services in Western Kenya. This low screening rate aligned with the reports that only 16% of Kenyan Women accessed cervical cancer screening which was often due to lack of awareness, stigma or health access (Nwabichie et al., 2017; Jedy-Agba et al., 2020). This gap in screening rates contributes to the high number of deaths rates from cervical cancer in the region (Tadesse, 2015; Momenimovahed et al, 2023). This study's use of the machine leaning model technique to predict cervical cancer risk is an approach that is promising to identify women who need screening more urgently, especially where resources are scare like in Western Kenya (Al-Naggar, 2022).

5.1.2 Machine Learning Model Performance

The first hypothesis (Ha1: There is a significant difference among machine learning models in detecting and classifying cervical cancer in terms of accuracy, sensitivity, and

specificity) was tested by comparing five different machine learning models: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) (Table 9). Overall, the random forest model performed best as it correctly identified 94% of the cases (accuracy = 94.33%) and ruled out cervical cancer in 98% of the negative cases (specificity=98.37%). The random forest model, however, only detected 20% of the actual cancer cases (sensitivity=20.00%), likely because only 6% of the women involved in the study had cervical cancer which created an imbalanced dataset. In contrast, the logistic regression model was better at finding the true cancer cases (sensitivity = 70.00%) but it was less accurate at confirming negative cases (specificity=83.15%).

These differences in the models ability to correctly predict cancer cases and non-cancer cases confirmed the first hypothesis, showing that the models have unique strengths. The random forest's high accuracy and specificity makes it more ideal for confirming respondents who do not have cervical cancer while the logistic regression's ability to detect more cancer cases makes it suitable for initial screening (Breiman, 2001; Chang & Lin, 2011). The ROC curve showed that the logistic regression model and the Support Vector Machine model were best at distinguishing cancer from non-cancerous cases (AUC=0.76). This was followed by the by Random Forest (AUC=0.71) and the ANN model which performed poorly (AUC=0.46) this was due to challenges in handling the imbalanced data, even with techniques to balance it (He et al., 2016; Mahanama, 2020). The results from this study shows that simpler models like the Logistic Regression can sometimes outperform even more complex models like ANN in medical settings where data is limited (Song et al., 2017; Shailaja et al., 2018).

5.1.3 Comparison with Existing Screening Methods

The second hypothesis (Ha2: The developed machine learning model will be significantly better in detecting and classifying cervical cancer than existing screening methods) was evaluated by comparing the Logistic Regression model (the best model for initial screening due to high sensitivity compared to others) and the Random Forest model (the best for confirmatory testing due to high specificity compared to other fitted models) with conventional methods such as Pap smear, HPV testing, and liquid-based cytology.

Logistic regression model for initial screening: The logistic regression model's sensitivity of 70.00% indicated that it correctly identified 70% of the women with cervical cancer. This was lower than the HPV test which has the ability to detect over 90% of the cases (often ranging between 96.7% and 100%) due to its ability to identify HPV infections linked to cervical cancer (Haile, 2019; ESTAMPA study, 2023; Arbyn et al., 2015). The Pap smear sensitivity ranges from 60% to 95%. This overlaps with the logistic regression model but it can miss due to poor sample quality of human error (Song et al., 2017; Bora et al., 2017). The Liquid based cytology has a sensitivity ranging between 70 % and 90%. Liquid cytology performs similarly but it requires specialized equipment which are not widely available in most parts of Western Kenya (Masenya, 2011; Kumaresan, 2021). Based on these findings, the logistic regression sensitivity is competitive with liquid based cytology but it falls short of HPV testing, making it a viable but not a superior option for the initial screening in resource-limited settings (WHO, 2021).

Random forest for confirmatory Testing: The random forest model's specificity was 98.37% which means that it correctly identified 98% of the women in western Kenya without cervical cancer. Random forest outperformed Pap smear (>90%), the HPV testing

(has a specificity ranging between 70 and 80%) and liquid based cytology (has a specificity ranging between 70 and 90%). The high specificity reduced the false positives which minimizes the unnecessary follow up tests and patients anxiety which is really important in low-resource regions (Jedy-Agba et al., 2020; Petersen et al., 2022). The random forest accuracy of 94.33% is also high than many reported accuracies from traditional methods though the accuracy alone is less informative given the imbalanced nature of the dataset used (Schölkopf & Smola, 2002).

These findings supported the second hypothesis partially. The random forest's exceptional specificity makes it better than the methods existing for confirming the negative cancer cases which offers a cost-effective way of streamlining screening programs (Creswell & Sheikh, 2013).

The logistic regression sensitivity was however not superior to the HPV testing and it was comparable to the Pap smear which indicates that it was not a definite improvement on the primary screening methods. The findings from this study shows that combining the use of the Logistic Regression to identify high-risk women and the Random Forest to confirm negative cases could enhance screening efficiency by leveraging on the strengths of both of these models and complementing traditional methods (Siddique & Chow, 2021; Al-Naggar, 2022).

5.1.4 Influential Risk Factor

The third hypothesis (Ha3: There is at least one influential risk factor for cervical cancer screening) was confirmed by the feature importance analysis (Table 4.9). The Pap smear results were the best predictor of cervical cancer in most models which reflects its role in detecting cervical abnormalities (Bora et al., 2017; Yang et al., 2018). The use of

hormonal contraceptives had also a significant impact on cervical cancer. These finding is consistent with studies linking long term use of contraceptives to higher cervical risk cases (Plummer et al., 2016; LaVigne et al., 2017). The models also flagged HIV as a notable risk factor which aligns with its known impact in sub-Saharan Africa where HIV increase the susceptibility of cervical cancer (Arbyn et al., 2020; Matenge & Mash, 2018).

Age was a moderately important factor since the study showed that older women were at a much higher risk of having cervical cancer compared to younger women due to prolonged HPV exposure (World Health Organization, 2022; Torre et al., 2015). The number of sexual partners showed a weak effect but the link between sexual partners and HPV transmission merits future study (Plummer et al., 2016; Walboomers et al., 1999). These risk factors flagged by the different models fitted in this study highlight the need for targeted screening that is based on demographic and clinical profiles (Al-Naggar, 2022).

5.1.5 Implications for Cervical Cancer Screening

The findings from this study suggested a two-step screening strategy to be used in Western Kenya. A logistic regression model to be used to identify women at high risk of cervical cancer which will ensure more cancer cases are caught early. Random forest can be used to confirm the negative cancer cases reducing the unnecessary costs and rates. This approach can address the low screening uptake and high mortality rate that arises due to late diagnosis of the disease (Nwabichie et al., 2017; Tadesse, 2015).

The public health campaigns should educate women on the risks such as hormonal contraceptives, early sexual debut, and HIV. They should also encourage regular Pap smear and HPV vaccination.

The machine learning should be integrated into existing screening programs so that they can optimize the limited resources and limit type 1 and type 2 error during screening, aligning with global efforts to reduce cervical cancer disparities (WHO, 2021; Petersen et al., 2022).

5.1.6 Limitations

The study faced some challenges that affects its findings:

- There was low number if cervical cancer cases (6%) which made it hard for models like SVM, Random forest and ANN to detect true cases despite the various analysis functions to balance the data.
- The results are based on one dataset collected from Western Kenya which may not apply elsewhere.
- There was presence of missing values which arise from respondents not disclosing certain information which could have influenced the results even with data imputation method.
- The study also relied on published data for traditional screening method rather than doing direct testing. This limited the strength of comparison.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.0 CONCLUSION

Machine learning models were developed in this study to detect cervical cancer in Western Kenya, which contributes to the efforts aimed at reducing cervical cancer burden in the low-income region of Western. The results partially confirmed the second hypothesis, as it was evident that the random Forest model excelled at confirming negative cases, surpassing the traditional methods like Pap smear and HPV testing, while Logistic Regression was better at detecting cancer cases, though not outperforming the HPV testing method. The models showed different distinct strengths, and the key risk factors, like Pap smear results, HPV results, use of hormonal contraceptives, age, and age at sexual debut, were identified. The findings from this study support the use of a combined screening approach to improve early detection of cervical cancer and also to address the disparities in cervical cancer outcomes.

6.1 RECOMMENDATIONS

- A two-step screening approach should be used, where logistic regression is used for initial screening to identify high-risk women and random forest for confirmatory tests to reduce the false positives.
- Focus on high-risk groups: Screening to be prioritized for HIV positive women, with early sexual activity, who are on hormonal contraceptives and are elderly.

- Education Campaigns: Awareness should be raised about cervical cancer risk factors such as hormonal contraceptives, early sexual activity and higher number of sexual partners and promotion of screening and vaccination.

6.2 FUTURE RESEARCH

Future study should focus on testing the logistic regression and random forest model on diverse Kenyan setting and comparing the results directly with Pap smear and HPV testing.

Using an even larger sample size to determine whether the performance of the models will improve or remain the same.

REFERENCE

Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database, 2020, baaa010.

Akinyemiju, T. F. (2012). Socio-economic and health access determinants of breast and cervical cancer screening in low-income countries: analysis of the World Health Survey. PloS one, 7(11), e48834.

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet Global Health, 8(2), e191-e203.

Barlow, H. B. (1989). Unsupervised learning. Neural computation, 1(3), 295-311.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

Berek, J. S., & Berek, D. L. (2020). Berek & Novak's gynecology. Wolters Kluwer.

Bora, K., Chowdhury, M., Mahanta, L. B., Kundu, M. K., & Das, A. K. (2017). Automated classification of Pap smear images to detect cervical dysplasia. Computer Methods and Programs in Biomedicine, 138, 31–47.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Bruni, L., Albero, G., Serrano, B., Mena, M., Gómez, D., Muñoz, J., Bosch, F. X., & de Sanjosé, S. (2018). Human Papillomavirus and Related Diseases in Kenya. ICO/IARC Information Centre on HPV and Cancer (HPV Information Centre).

Burt, A. A., Nakisige, C., Ntege, E., Atukunda, R., Kakande, P., Mutyaba, T., Anderson, M., Gage, J. C., Castle, P. E., & Wentzensen, N. (2021). HPV-based screening and treatment of high-grade cervical disease in rural western Uganda: A prospective cohort study. American Journal of Obstetrics & Gynecology, 225(6), S23–S24

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

Cancer Research UK. (2015, March 17). Signs and symptoms of cancer. Cancer Research UK.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Cohen, P. A., Jhingran, A., Oaknin, A., Denny, L. (2019). Cervical cancer. Lancet, 393(10167), 169–182. https://doi.org/10.1016/S0140-6736(18)32470-X.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In Machine learning techniques for multimedia: case studies on organization and retrieval (pp. 21-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

Dunteman, G. H. (1989). Principal component analysis. Sage.

Feltovich, P. J., Ford, K. M., & Hoffman, R. R. (1997). Expertise in context: human and machine. Aaai Press; Cambridge, Mass.

Fernandes, K., Cardoso, J. S., & Fernandes, J. (2018). Transfer learning with partial observability applied to cervical cancer screening. Pattern Recognition Letters, 111, 23-30.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, No. 2018). Cham: Springer.

Field, A. (2005). Discovering statistics using SPSS (3rd ed.). London: Sage.

Fink, A. (2009). How to conduct surveys: A step-by-step guide (4th ed.). Thousand Oaks, CA: Sage.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. International Journal of Intelligent Technologies and Applied Statistics, 11(2), 105-111.

Global Cancer Observatory. (2018). Iarc.fr.

Haile, E. L. (2019). HPV Testing on Vaginal/Cervical Nurse Assisted Self-Samples Versus Clinician-Taken Specimens and EHE HPV Prevalence, in Adama Town, Ethiopia. Ergonomics International Journal, 3(3).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). New York, NY: Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

International Agency for Research on Cancer & World Health Organization. (2020). Globocan 2020. Available at gco.iarc.fr. Accessed February 3, 2021.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks, A tutorial. Computer, 29(3), 31–44.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kashyap, D., Garg, V. K., Tuli, H. S., Yerer, M. B., Sak, K., Sharma, A. K., Kumar, M., Aggarwal, V., & Sandhu, S. S. (2019). Fisetin and quercetin: Promising flavonoids with chemopreventive potential. Biomolecules, 9(5), Article 174.

Kenya National Bureau of Statistics. (2023). Kenya National Bureau of Statistics. https://www.knbs.or.ke/

Kieffer, J. (1994). Elements of information theory (Thomas M. Cover and Joy A. Thomas). SIAM Review, 36(3), 509-511.

Kumaresan, D. (2021). How Liquid Based Cytology Surpasses Conventional Cytology - A Review Article. Journal of Pharmaceutical Research International, 469–471. https://doi.org/10.9734/jpri/2021/v33i59a34293.

LaVigne, A. W., Triedman, S. A., Randall, T. C., Trimble, E. L., & Viswanathan, A. N. (2017). Cervical cancer in low and middle income countries: Addressing barriers to radiotherapy delivery. Gynecologic Oncology Reports, 22, 16–20. https://doi.org/10.1016/j.gore.2017.08.004.

Lilhore, U. K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A. D., Elmannai, H., ... & Hamdi, M. (2022). Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. Computational and Mathematical Methods in Medicine, 2022(1), 4688327.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. Annals of statistics, 42(2), 413.

Mahanama, S. (2020). Introduction to Artificial Neural Networks (ANN). Analytics Vidhya. https://medium.com/analytics-vidhya/introduction-to-artificial-neural-networks-ann-3109578d61ab.

Martin, C. M., Astbury, K., McEvoy, L., Toole, S., Sheils, O., & Leary, J. J. (2009). Gene expression profiling in cervical cancer: Identification of novel markers for disease diagnosis and therapy. In Inflammation and Cancer; Springer: Berlin, Germany, Volume 511, pp. 333–359.

Masenya, M. (2011). Liquid based cytology. Obstetrics and Gynaecology Forum, 21(3). https://doi.org/10.4314/ogf.v21i3.69484.

Matenge, T. G., & Mash, B. (2018). Barriers to accessing cervical cancer screening among HIV positive women in Kgatleng district, Botswana: A qualitative study. PLoS One, 13(10), e0205425. https://doi.org/10.1371/journal.pone.0205425.

Momenimovahed, Z., Mazidimoradi, A., Maroofi, P., Allahqoli, L., Salehiniya, H., & Alkatout, I. (2023). Global, regional and national burden, incidence, and mortality of cervical cancer. Cancer reports, 6(3), e1756.

National Cancer Institute. (2021, October 11). What Is Cancer? National Cancer Institute; Cancer.gov.

Nocedal, J., & Wright, S. J. (2006). Large-scale unconstrained optimization. Numerical optimization, 164-192.

Nwabichie, C. C., Rosliza, A. M., & Suriani, I. (2017). Global Burden of Cervical Cancer: A Literature Review. BMC Public Health, 17(1), 1157.

Oh, K., Kang, H. M., Leem, D., Lee, H., Seo, K. Y., & Yoon, S. (2021). Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-81539-3.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

Pitts, N. B., & Richards, D. (2009). Personalized treatment planning. Detection, Assessment, Diagnosis and Monitoring of Caries, 21, 128-143.

Purnami, S., Khasanah, P., Sumartini, S., Chosuvivatwong, V., & Sriplung, H. (2016). Cervical cancer survival prediction using hybrid of SMOTE, CART and smooth support vector machine. AIP Conf. Proc., 1723, 030017.

Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., & Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: A synthetic analysis. The Lancet Global Health, 4(9), e609–e616.

Raza, K., & Singh, N. K. (2021). A tour of unsupervised deep learning for medical image analysis. Current Medical Imaging, 17(9), 1059-1077.

Republic of Kenya, Ministry of Health. (2017). National Cancer Control Strategy 2017–2022.

Rothman, K. J., Sander Greenland, & Lash, T. L. (2019). Modern epidemiology. Wolters Kluwer Health / Lippincott Williams & Wilkins.

Schork, N. J. (2019). Artificial Intelligence and Personalized Medicine. Precision Medicine in Cancer Therapy, 265–283. https://doi.org/10.1007/978-3-030-16391-4_11.

Sfeir, J. G., Kittah, N. E. N., Tamhane, S. U., Jasim, S., Chemaitilly, W., Cohen, L. E., & Murad, M. H. (2018). Diagnosis of GH Deficiency as a Late Effect of Radiotherapy in Survivors of Childhood Cancers. The Journal of Clinical Endocrinology & Metabolism, 103(8), 2785–2793. https://doi.org/10.1210/jc.2018-01204.

Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018). Machine learning in healthcare: A review. In 2018 Second international conference on electronics, communication and aerospace technology (ICECA) (pp. 910-914). IEEE.

Sharma, S. (2017). Activation Functions in Neural Networks. Towards Data Science. https://towardsdatascience.com/activation-functions-neural-networks.

Siddique, S., & Chow, J. C. (2021). Machine learning in healthcare communication. Encyclopedia, 1(1), 220-239.

Song, Y., Tan, E.-L., Jiang, X., Cheng, J.-Z., Ni, D., Chen, S., Lei, B., & Wang, T. (2017). Accurate Cervical Cell Segmentation from Overlapping Clumps in Pap Smear Images. IEEE Transactions on Medical Imaging, 36(1), 288–300.

Tadesse, S. K. (2015). Socio-economic and cultural vulnerabilities to cervical cancer and challenges faced by patients attending care at Tikur Anbessa Hospital: A cross sectional and qualitative study. BMC Women's Health, 15, Article 75.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians, 65(2), 87–108. https://doi.org/10.3322/caac.21262.

Walboomers, J. M. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., Snijders, P. J. F., Peto, J., Meijer, C. J. L. M., & Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. The Journal of Pathology, 189(1), 12–19.

World Health Organization. (2020). A Global Strategy for Elimination of Cervical Cancer. https://www.who.int/activities/a-global-strategy-for-elimination-of-cervical-cancer.

World Health Organization. (2022, February 22). Cervical Cancer. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cervical-cancer.

Yang, X., Da, M., Zhang, W., Qi, Q., Zhang, C., & Han, S. (2018). Role of lactobacillus in cervical cancer. Cancer Management and Research, 10, 1219–1229. https://doi.org/10.2147/CMAR.S165764.

APPENDICES

APPENDIX I: CODES

```
# Step 1: Import the pandas library

import pandas as pd

# Step 2: Use the pandas.read_csv() function to read the CSV file

# Replace 'data.csv' with the actual name of your CSV file if different

df = pd.read_csv('C:\\Users\\Administrator\\ cervical.csv')

# Step 3: Display the data using the head() method (optional)

df.head()

# Checking data types

data_types = df.dtypes

# Display the data types

print(data_types)

# Function to create frequency and percentage table

def freq_percent_table(column):

    freq = pd.crosstab(index=df[column], columns='Frequency')

    percent = pd.crosstab(index=df[column], columns='Percentage', normalize=True) * 100

    freq_percent = freq.join(percent)

    return freq_percent



# Frequency and percentage table for Smokes

smokes_table = freq_percent_table('Smokes')

print("Frequency and Percentage Table for Smokes:")

print(smokes_table)
```

```
print("\n")

# Frequency and percentage table for Age_Category

age_table = freq_percent_table('Age_Category')

print("Frequency and Percentage Table for Age_Category:")

print(age_table)

print("\n")

# Frequency and percentage table for Hormonal Contraceptives

hormonal_table = freq_percent_table('Hormonal Contraceptives')

print("Frequency and Percentage Table for Hormonal Contraceptives:")

print(hormonal_table)
print("\n")

# Frequency and percentage table for Pap smear

citology_table = freq_percent_table('Citology')

print("Frequency and Percentage Table for Citology:")

print(citology_table)

print("\n")

# Frequency and percentage table for Biopsy

biopsy_table = freq_percent_table('Biopsy')

print("Frequency and Percentage Table for Biopsy:")

print(biopsy_table)

print("\n")

# Using errors='coerce' to convert non-numeric values to NaN

df['Number of sexual partners'] = pd.to_numeric(df['Number of sexual partners'],
errors='coerce')

df['First sexual intercourse'] = pd.to_numeric(df['First sexual intercourse'],
errors='coerce')
```

```
df['Num of pregnancies'] = pd.to_numeric(df['Num of pregnancies'], errors='coerce')

df['Dx:Cancer'] = pd.to_numeric(df['Dx:Cancer'], errors='coerce')

df['Num of pregnancies'] = pd.to_numeric(df['Num of pregnancies'], errors='coerce')

df['Smokes (years)'] = pd.to_numeric(df['Smokes (years)'], errors='coerce')

# Calculate mean and standard deviation for each variable

mean_sexual_partners = df['Number of sexual partners'].mean()

std_sexual_partners = df['Number of sexual partners'].std()

mean_first_intercourse = df['First sexual intercourse'].mean()

std_first_intercourse = df['First sexual intercourse'].std()

mean_pregnancies = df['Num of pregnancies'].mean()

std_pregnancies = df['Num of pregnancies'].std()

# Display the results

print("Number of sexual partners:")

print(f"Mean: {mean_sexual_partners:.2f}, Standard Deviation: {std_sexual_partners:.2f}")

print("\n")

print("First sexual intercourse:")

print(f"Mean: {mean_first_intercourse:.2f}, Standard Deviation:
{std_first_intercourse:.2f}")

print("\n")

print("Num of pregnancies:")

print(f"Mean: {mean_pregnancies:.2f}, Standard Deviation: {std_pregnancies:.2f}")

 # Machine learning model

def calculate_specificity(y_true, y_pred):

    tn, fp, fn, tp = confusion_matrix(y_true, y_pred).ravel()

    specificity = tn / (tn + fp) if (tn + fp) > 0 else 0
```

```
    return specificity

# Loading the packages

import numpy as np

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy_score, confusion_matrix, precision_score,
recall_score

from sklearn.impute import KNNImputer

# Preprocess the data

# Replace '?' with NaN

new_df1 = new_df1.replace('?', np.nan)

# Convert object columns to numeric where appropriate

object_cols = new_df1.select_dtypes(include='object').columns

for col in object_cols:

    new_df1[col] = pd.to_numeric(new_df1[col], errors='coerce')

# Check for remaining non-numeric columns and encode them if needed

for col in new_df1.select_dtypes(include='object').columns:

    new_df1[col] = new_df1[col].astype('category').cat.codes

# Apply KNN imputation to handle NaN values

imputer = KNNImputer(n_neighbors=5)  # Using 5 neighbors as a reasonable default

new_df1_imputed = pd.DataFrame(imputer.fit_transform(new_df1), columns=new_df1.columns)

# Verify no missing values remain

print("Missing Values After KNN Imputation:")

print(new_df1_imputed.isnull().sum())
```

```
# Verify data types

print("\nData Types After Preprocessing:")

print(new_df1_imputed.dtypes)

# Splitting the data

X = new_df1_imputed.drop('Biopsy', axis=1)

y = new_df1_imputed['Biopsy']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Helper function

def calculate_specificity(y_true, y_pred):

    tn, fp, fn, tp = confusion_matrix(y_true, y_pred).ravel()

    specificity = tn / (tn + fp) if (tn + fp) > 0 else 0

    return specificity

# Decision tree

# Initialize and train Decision Tree with class weighting

dt_classifier = DecisionTreeClassifier(random_state=42, max_depth=5,
class_weight='balanced')

dt_classifier.fit(X_train, y_train)

y_pred_dt = dt_classifier.predict(X_test)

# Evaluate

accuracy_dt = accuracy_score(y_test, y_pred_dt)

precision_dt = precision_score(y_test, y_pred_dt, pos_label=1)

recall_dt = recall_score(y_test, y_pred_dt, pos_label=1)

specificity_dt = calculate_specificity(y_test, y_pred_dt)

conf_matrix_dt = confusion_matrix(y_test, y_pred_dt)

print("Decision Tree Classifier Performance (Balanced):")
```

```
print(f"Accuracy: {accuracy_dt:.4f}")

print(f"Precision (Positive Class): {precision_dt:.4f}")

print(f"Sensitivity (Recall, Positive Class): {recall_dt:.4f}")

print(f"Specificity (Negative Class): {specificity_dt:.4f}")

print("\nConfusion Matrix:")

print(conf_matrix_dt)

# Feature Importance for Decision Tree

# Extract feature importances

feature_importances = dt_classifier.feature_importances_

features = X_train.columns

# Create a DataFrame for better visualization

importance_df = pd.DataFrame({

    'Feature': features,

    'Importance': feature_importances

}).sort_values(by='Importance', ascending=False)

# Display feature importances

print("\nDecision Tree Feature Importance:")

print(importance_df)

# Random forest

from sklearn.ensemble import RandomForestClassifier

# Initialize and train Random Forest with class weighting

rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42, max_depth=5,
class_weight='balanced')

rf_classifier.fit(X_train, y_train)

y_pred_rf = rf_classifier.predict(X_test)
```

```
# Evaluate

accuracy_rf = accuracy_score(y_test, y_pred_rf)

precision_rf = precision_score(y_test, y_pred_rf, pos_label=1)

recall_rf = recall_score(y_test, y_pred_rf, pos_label=1)

specificity_rf = calculate_specificity(y_test, y_pred_rf)

conf_matrix_rf = confusion_matrix(y_test, y_pred_rf)

print("Random Forest Classifier Performance (Balanced):")

print(f"Accuracy: {accuracy_rf:.4f}")

print(f"Precision (Positive Class): {precision_rf:.4f}")

print(f"Sensitivity (Recall, Positive Class): {recall_rf:.4f}")

print(f"Specificity (Negative Class): {specificity_rf:.4f}")

print("\nConfusion Matrix:")

print(conf_matrix_rf)

 # Extract feature importances

feature_importances = rf_classifier.feature_importances_

features = X_train.columns

# Create a DataFrame for better visualization

importance_df = pd.DataFrame({

    'Feature': features,

    'Importance': feature_importances

}).sort_values(by='Importance', ascending=False)

# Display feature importances

print("\nRandom Forest Feature Importance:")
```

```
print(importance_df)

# Logit model

from sklearn.linear_model import LogisticRegression

# Initialize and train Logistic Regression with class weighting

logit_classifier = LogisticRegression(random_state=42, max_iter=1000,
class_weight='balanced')

logit_classifier.fit(X_train, y_train)

y_pred_logit = logit_classifier.predict(X_test)



# Evaluate

accuracy_logit = accuracy_score(y_test, y_pred_logit)

precision_logit = precision_score(y_test, y_pred_logit, pos_label=1)

recall_logit = recall_score(y_test, y_pred_logit, pos_label=1)

specificity_logit = calculate_specificity(y_test, y_pred_logit)

conf_matrix_logit = confusion_matrix(y_test, y_pred_logit)

print("Logistic Regression Classifier Performance (Balanced):")

print(f"Accuracy: {accuracy_logit:.4f}")

print(f"Precision (Positive Class): {precision_logit:.4f}")

print(f"Sensitivity (Recall, Positive Class): {recall_logit:.4f}")

print(f"Specificity (Negative Class): {specificity_logit:.4f}")

print("\nConfusion Matrix:")

print(conf_matrix_logit)

 # SVM

 from sklearn.svm import SVC

# Initialize and train SVM with class weighting
```

```
svm_classifier = SVC(kernel='linear', random_state=42, probability=True,
class_weight='balanced')

svm_classifier.fit(X_train, y_train)

y_pred_svm = svm_classifier.predict(X_test)

# Evaluate

accuracy_svm = accuracy_score(y_test, y_pred_svm)

precision_svm = precision_score(y_test, y_pred_svm, pos_label=1)

recall_svm = recall_score(y_test, y_pred_svm, pos_label=1)

specificity_svm = calculate_specificity(y_test, y_pred_svm)

conf_matrix_svm = confusion_matrix(y_test, y_pred_svm)

print("Support Vector Machine Classifier Performance (Balanced):")

print(f"Accuracy: {accuracy_svm:.4f}")

print(f"Precision (Positive Class): {precision_svm:.4f}")

print(f"Sensitivity (Recall, Positive Class): {recall_svm:.4f}")

print(f"Specificity (Negative Class): {specificity_svm:.4f}")

print("\nConfusion Matrix:")

print(conf_matrix_svm)

 # Feature Importance for Linear SVM
# Extract feature coefficients (weights)
feature_coefficients = svm_classifier.coef_[0]  # coef_ returns a 2D array; take first
row for binary classification
features = X_train.columns

# Create a DataFrame for better visualization
# Use absolute values of coefficients to measure importance (direction doesn't matter for
importance)
importance_df = pd.DataFrame({
    'Feature': features,
    'Coefficient': feature_coefficients,
    'Importance': np.abs(feature_coefficients)  # Absolute value for ranking
}).sort_values(by='Importance', ascending=False)

# Display feature importances
```

```
print("\nSupport Vector Machine Feature Importance (Based on Coefficients):")
print(importance_df)

# ANN
from sklearn.neural_network import MLPClassifier
from imblearn.over_sampling import SMOTE
from sklearn.metrics import accuracy_score, precision_score, recall_score,
confusion_matrix
from sklearn.inspection import permutation_importance
import pandas as pd
import matplotlib.pyplot as plt

# Apply SMOTE to balance the training data
smote = SMOTE(random_state=42)
X_train_balanced, y_train_balanced = smote.fit_resample(X_train, y_train)

# Initialize and train ANN on balanced data
ann_classifier = MLPClassifier(hidden_layer_sizes=(100, 50), max_iter=500,
random_state=42,
                               activation='relu', solver='adam')
ann_classifier.fit(X_train_balanced, y_train_balanced)
y_pred_ann = ann_classifier.predict(X_test)

# Evaluate
accuracy_ann = accuracy_score(y_test, y_pred_ann)
precision_ann = precision_score(y_test, y_pred_ann, pos_label=1)
recall_ann = recall_score(y_test, y_pred_ann, pos_label=1)
specificity_ann = calculate_specificity(y_test, y_pred_ann)
conf_matrix_ann = confusion_matrix(y_test, y_pred_ann)

print("Artificial Neural Network Classifier Performance (SMOTE Balanced):")
print(f"Accuracy: {accuracy_ann:.4f}")
print(f"Precision (Positive Class): {precision_ann:.4f}")
print(f"Sensitivity (Recall, Positive Class): {recall_ann:.4f}")
print(f"Specificity (Negative Class): {specificity_ann:.4f}")
print("\nConfusion Matrix:")
print(conf_matrix_ann)

# Feature Importance for ANN using Permutation Importance
# Compute permutation importance on the test set
perm_importance = permutation_importance(ann_classifier, X_test, y_test,
                                         n_repeats=10, random_state=42,
scoring='accuracy')

# Extract feature importances and standard deviations
feature_importances = perm_importance.importances_mean
features = X_train.columns
```

```
# Create a DataFrame for better visualization
importance_df = pd.DataFrame({
    'Feature': features,
    'Importance': feature_importances,
    'Std': perm_importance.importances_std
}).sort_values(by='Importance', ascending=False)

# Display feature importances
print("\nArtificial Neural Network Feature Importance (Permutation Importance):")
print(importance_df)

# ROC
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, roc_auc_score

# All models are trained with imbalance adjustments
models = {
    'Decision Tree': dt_classifier,         # class_weight='balanced'
    'Random Forest': rf_classifier,         # class_weight='balanced'
    'Logistic Regression': logit_classifier, # class_weight='balanced'
    'SVM': svm_classifier,                  # class_weight='balanced'
    'ANN': ann_classifier                   # SMOTE-balanced
}

# Dictionary to store ROC data
roc_data = {}

# Calculate ROC curve and AUC for each model
for model_name, model in models.items():
    # Get probability scores for the positive class (Biopsy = 1)
    y_prob = model.predict_proba(X_test)[:, 1]
    # Compute ROC curve
    fpr, tpr, _ = roc_curve(y_test, y_prob, pos_label=1)
    # Compute AUC
    auc = roc_auc_score(y_test, y_prob)
    roc_data[model_name] = (fpr, tpr, auc)

# Plot all ROC curves in one figure
plt.figure(figsize=(10, 8))
for model_name, (fpr, tpr, auc) in roc_data.items():
    plt.plot(fpr, tpr, label=f'{model_name} (AUC = {auc:.2f})')

# Plot the random guess line
plt.plot([0, 1], [0, 1], 'k--', label='Random Guess (AUC = 0.50)')

# Customize the plot
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate (Sensitivity)')
```

```
plt.title('ROC Curves for Cervical Cancer Prediction Models (Balanced)')
plt.legend(loc='lower right')
plt.grid(True)
plt.tight_layout()

# Show the plot
plt.show()

# Print AUC scores separately for reference
print("ROC AUC Scores:")
for model_name, (_, _, auc) in roc_data.items():
    print(f"{model_name}: {auc:.4f}")

import matplotlib.pyplot as plt
from sklearn.tree import plot_tree

# Plot the Decision Tree
plt.figure(figsize=(20, 10))  # Large size for readability
plot_tree(dt_classifier,
          feature_names=X_train.columns,
          class_names=['No', 'Yes'],
          filled=True,
          rounded=True,
          fontsize=10,
          max_depth=3)  # Limit depth for clarity; remove or adjust as needed
plt.title('Decision Tree for Cervical Cancer Prediction (Balanced)')
plt.show()
R code for data visualization
# Data importation
df = read.csv('C:\\Users\\Administrator\\Downloads\\cervical.csv')

# View the data
head(df)
# Figure showing Cytology results distribution
# Load required library
library(ggplot2)
# Create the plot
ggplot(df, aes(x = factor(Citology))) +
  geom_bar(fill = "blue") +
  labs(title = "Pap smear distribution",
       x = "Pap smear Result",
       y = "Count") +
  scale_x_discrete(labels = c("0" = "Negative", "1" = "Positive"))+
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.title = element_text(size = 12)
  )+
  theme_minimal()
```

```
# Figure showing biopsy results distribution
# Plot biopsy results distribution
ggplot(df, aes(x = factor(Biopsy))) +
  geom_bar(fill = "blue", width = 0.7) +
  labs(title = "Biopsy Results Distribution",
       x = "Biopsy Result",
       y = "Count") +
  theme_minimal() +
  scale_x_discrete(labels = c("0" = "Negative", "1" = "Positive")) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.title = element_text(size = 12)
  )

df$Num.of.pregnancies<-as.numeric(df$Num.of.pregnancies)
df$Number.of.sexual.partners<-as.numeric(df$Number.of.sexual.partners)
library(dplyr)

# Calculate mean and SD
summary_stats <- df %>%
  group_by(Biopsy) %>%
  summarise(
    Mean_Partners = mean(Number.of.sexual.partners, na.rm = TRUE),
    SD = sd(Number.of.sexual.partners, na.rm = TRUE)
  )
# Plot
ggplot(summary_stats, aes(x = factor(Biopsy), y = Mean_Partners)) +
  geom_bar(stat = "identity", fill = c("#4E79A7", "#F28E2B"), width = 0.6) +
  geom_errorbar(
    aes(ymin = Mean_Partners - SD, ymax = Mean_Partners + SD),
    width = 0.2,
    color = "black"
  ) +
  labs(
    title = "Mean Number of Sexual Partners by Biopsy Result",
    x = "Biopsy Result",
    y = "Mean Number of Sexual Partners",
    caption = "Error bars represent ±1 standard deviation"
  ) +
  scale_x_discrete(labels = c("Negative (0)", "Positive (1)")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.title = element_text(size = 12),
    panel.grid.major.x = element_blank()
  )
# Convert 'First.sexual.intercourse' to numeric (handle '?' as NA)
```

```
df_clean <- df %>%
  mutate(
    First.sexual.intercourse = as.numeric(ifelse(First.sexual.intercourse == "?", NA,
First.sexual.intercourse))
  ) %>%
  filter(!is.na(First.sexual.intercourse), !is.na(Biopsy))
```

APPENDIX II: NACOSTI

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 890285

Date of Issue: 18/June/2024

RESEARCH LICENSE

This is to Certify that Mr.. John Murere Flavian of University of Eldoret, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in on the topic: MACHINE LEARNING BASED CERVICAL CANCER DISEASE DETECTION AND CLASSIFICATION MODEL IN WESTERN KENYA for the period ending : 18/June/2025.

License No: NACOSTI/P/24/36733

890285

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013 (Rev. 2014)
Legal Notice No. 108: The Science, Technology and Innovation (Research Licensing) Regulations, 2014

The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was the established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of International treaties of which Kenya is a signatory to
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
    i. Endanger national security
    ii. Adversely affect the lives of Kenyans
    iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
    iv. Result in exploitation of intellectual property rights of communities in Kenya
    v. Adversely affect the environment
    vi. Adversely affect the rights of communities
    vii. Endanger public safety and national cohesion
    viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. The license any rights thereunder are non-transferable
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.
13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and Innovation(NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 0713788787, 0735404245
E-mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke

APPENDIX III: NACOSTI

MASINDE MULIRO UNIVERSITY OF SCIENCE AND TECHNOLOGY

Tel: 056-31375
Fax: 056-30153
E-mail: ierc@mmust.ac.ke
Website: www.mmust.ac.ke

P. O. Box 190,
50100.
Kakamega,
KENYA

Institutional Scientific and Ethics Review Committee (ISERC)

To: Mr. Murere J Flavian

Date: May 20th, 2024

Dear Mr.

RE: MACHINE LEARNING BASED CERVICAL CANCER DISEASE DETECTION AND CLASSIFICATION MODEL IN WESTERN KENYA.

This is to inform you that the ***Masinde Muliro University of Science and Technology Institutional Scientific and Ethics Review Committee (MMUST-ISERC)*** has reviewed and approved your above research proposal. Your application approval number is **MMUST/ ISERC/048/2024.** The approval covers for the period ***May 20th, 2024 to May 20th, 2025.***
This approval is subject to compliance with the following requirements;

i. Only approved documents including informed consents, study instruments, MTA will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by ***MMUST-ISERC.***
iii. Death and life threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to ***MMUST-ISERC*** within 72 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to ***MMUST-ISERC*** within 72 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to ***MMUST-ISERC.***

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke and also obtain other clearances needed

Yours Sincerely,

Prof. Gordon Nguka (PhD)
Chairperson, Institutional Scientific and Ethics Review Committee

Copy to:
- The Secretary, National Bio-Ethics Committee
- Vice Chancellor
- DVC (PR&I)

APPENDIX IV: SIMILARITY REPORT

University of Eldoret

University of Eldoret

Certificate of Plagiarism Check for Thesis

Author Name	Murere John Flavian SSCI/MAT/M/011/22
Course of Study	Type here...
Name of Guide	Type here...
Department	Type here...
Acceptable Maximum Limit	Type here...
Submitted By	titustoo@uoeld.ac.ke
Paper Title	MACHINE LEARNING BASED CERVICAL CANCER DETECTION MODEL IN WESTERN KENYA
Similarity	11%
Paper ID	4635952
Total Pages	108
Submission Date	2025-11-06 19:18:32

Signature of Student

Signature of Guide

University of Eldoret LIBRARIAN 07 NOV 2025 P.O. Box 1125, ELDORET - 30100

Head of the Department

University Librarian

Director of Post Graduate Studies

* This report has been generated by DrillBit Anti-Plagiarism Software