International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 9 Issue 6 June - 2025, Pages: 34-40

Detection and classification of cervical cancer disease among women using machine learning technique Model in Western Kenya.

JF Murere[1]*, S. Wangila[1], J. Koech[1]

1Department of Mathematics and Computer Sciences, University of Eldoret (UE) P.O Box 1125-30100, Eldoret, Kenya.

Corresponding Author: John Flavian Murere. flavianmurere2@gmail.com

Abstract: *Cervical cancer is the leading cause of cancer related deaths among Kenyan women, claiming approximately the lives of 3,200 women annually. This is primarily due to the low screening uptake (16%) and late diagnosis. The aim of this study was to develop a machine leaning based model to enhance early detection of cervical cancer in Western Kenya, a region in Kenya with limited healthcare resources. Demographic, reproductive, and clinical characteristics data were collected from 968 women across health facilities in western Kenya (MTRH and Kakamega Referral hospital) utilizing a cross sectional study design. The dataset was divided into training set (70%) and testing set (30%). The training set was used to develop the five machine learning model: Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN). The testing set was used to evaluate the models. The machine learning model were trained to classify the cervical cancer cases, addressing the class imbalances using class weighting method for SVM, decision tree, random forest and logit model and synthetic minority oversampling class technique (SMOTE) for ANN. The random forest model demonstrated the superior performance compared to the other four models as it achieved the highest accuracy (94.33%) and specificity (98.37%) making it to be highly effective at ruling out negative cases. It however had a sensitivity of 20% which indicated that it had challenges in detecting positive cases. The logistic regression model excelled in sensitivity (70%) making it suitable for initial screening. ANN model showed the lowest precision (10%). The findings from this study suggested that a two-step approach which combine both Logistic Regression for screening and Random Forest for confirmation of cervical cancer cases which will go a long way in improving early detection and reduce cervical cancer mortality in resource-constrained settings like Western Kenya.*

Keywords: Cervical Cancer, Mortality, Classification, Detection, Model, Sensitivity.

INTRODUCTION

1.0 Background information

Cervical cancer among the commonest females cancers and is the leading cause of mortality in many developing countries (Arbyn et al., 2020). Cancer is a medical condition where some of the cells in the body grow uncontrollably and then spread to other parts of the body (National Cancer Institute, 2021). Cervical cancer continues to remain a big health problem globally and it is the fourth common cancer that is known to affect women globally (Lilhore et al., 2022). Whereas western countries have significantly reduced cervical cancer deaths overtime, a big gap still exist in developing countries where its estimated that approximately 90% of the burden of mortality occur in these regions and the mortality rate has been estimated to be 18 times higher than the rates in developed countries (Akinyemiju, 2012). 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). Middle and low-income countries have an equal share of 83% of the world's cervical cancer burden. Nevertheless, the screening coverage is mealy 19%, different from high income countries with lower cervical cancer burden and higher screening

coverage of 63% ((Matenge & Mash, 2018)).In Sub-Saharan Africa, the mortality rates for cervical cancer remains among the highest globally. The government has made significant strides

to increase the screening uptake, but still, the number of deaths and cases continues to rise and is projected to rise by the Ministry of Health. Factors contributing to this disparity include access to detection, screening, and other healthcare services. Research shows that there are numerous potential risk factors that affect cervical cancer among women and this poses treatment prediction challenges. 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 and classification 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. This study aimed to leverage machine learning tools to develop a prognostic model to help clinicians, doctors, and healthcare practitioners detect cervical cancer early and assess the risk of a patient 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. To develop the most suitable machine learning technique for the detection and classification of the cervical cancer disease among women in Western Kenya.

MATERIAL AND METHODS

3.1 Research Methodology

This study employed a cross-sectional study design. The study looked back in time to collect information on the past exposure of the respondents who were involved in the study. Data for the cross-sectional in Western Kenya (Kakamega Referral Hospital, MTRH and Kitale district). The data was obtained from patients who had undergone cervical cancer testing and received their results. After the data had been collected, the data was divided into training set and testing set.

study design was obtained using both an open-ended and closed-ended questionnaire. The source of the dataset was the hospital and healthcare facilities that are locate

3.2 Model Development Flow

The development process started with preparing the data that was to be used for training and testing. 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. Feature engineering among other techniques was used during the training stage and the model trained be evaluated with the test set and their inference exported.

The developed models were put into test in this phase. The testing data was used to test the performance of the model on unseen data. The comparison between the models was done and the best model picked using proper instruments and metrics.

3.4 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.

3.4.1 Data Pre-processing Steps

The study checked for missing values, duplication, and other potential errors. In cases where the missing value was on the dependent variable (diagnosis of cervical cancer), the whole row was eliminated from the study.

However, if the missing value was on other independent variables, this study used an imputation method (KNN) to fill in the missing values.

3.4.2 Handling Class Imbalance

The synthetic minority oversampling technique (SMOTE) was used to balance the training data for the artificial neural network model. 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).

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).

3.5 Data Partitioning

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 dataset was

partitioned into training and testing sets following the 70-30 rule which is a widely accepted guideline (Gholamy et al., 2018).

3.6 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.

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. The oncologist, epidemiologist, and public health professionals are some of the domain experts that were consulted.

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.

3.6.1 Feature Selection Technique to Reduce Variability
This study used principal component analysis and linear discriminant analysis to reduce the dimension in the data before fitting a prognostic model.

3.7 The Model Development
This study aimed to develop a reliable and accurate detection and classification model for predicting cervical cancer occurrence 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.

3.7.1 The Logistic Regression Model

A logit regression model was used to model the relationship between predictors and the response variable. This model provided the impact of each particular predictor on the probability of developing cervical cancer.

3.7.2 The Random Forest Model
The study also used a random forest model. Random forest is a model that uses many individual decision trees that operate as an ensemble (Breiman, 2001).

3.7.3 The 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.

3.7.4 Artificial Neural Network
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.

3.8 The Model Testing Component

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.

3.9 Evaluation Metrics

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. Accuracy represented the proportion of the cervical cancer cases that were correctly classified. Precision represented the proportion of cervical cancer cases that were correctly predicted to be positive out of all cervical cancer predicted as positive. Recall represented the proportion of cervical cancer cases that were correctly classified to be positive out of all the actual positive cervical cancer cases

RESULTS

4.4.1 Comparative Analysis

Comparative analysis results are summarized in the table below:

Table 4.1: *Table showing 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 4.1 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 SVM machine 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.

4.4.2 Visualization of Model Performance

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

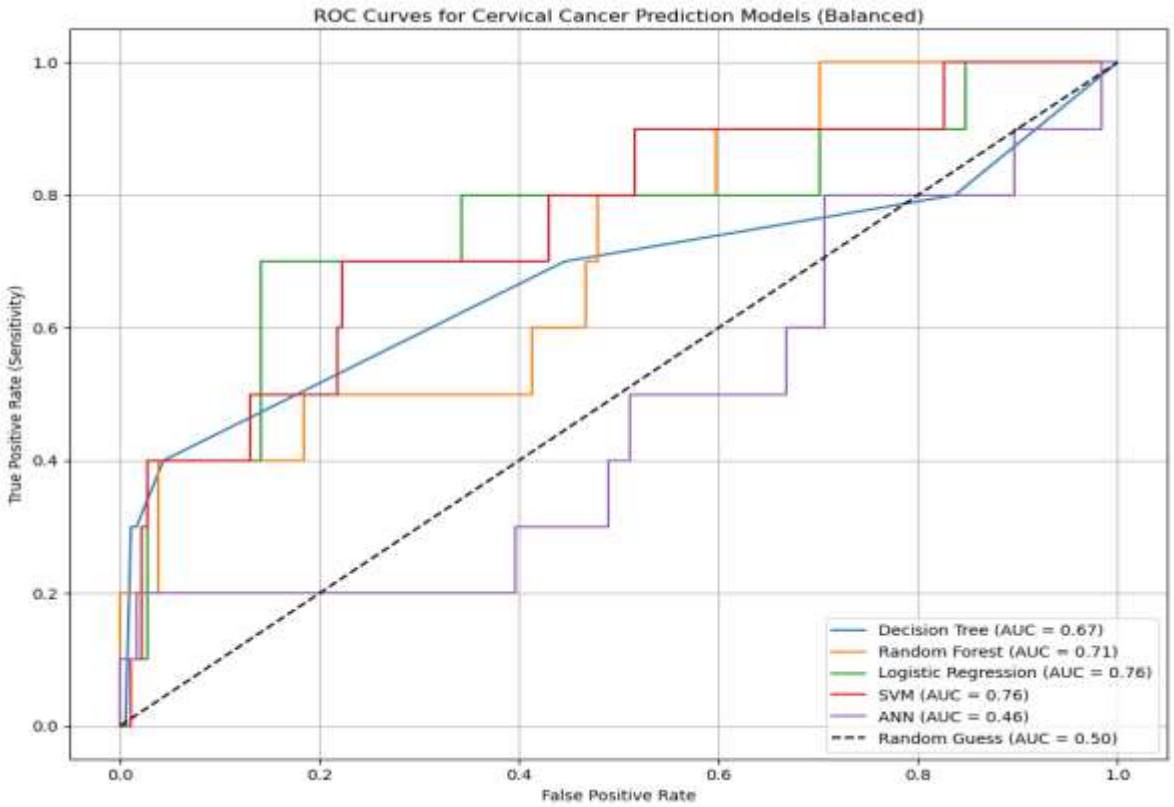

Figure 4.6: *ROC plot*

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.

DISCUSSION AND CONCLUSION

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. 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. 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 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; Cancer Atlas, 2019).

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). 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).

5.2 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 hypothesis as it was evident that the random Forest model excelled at

confirming negative cases while Logistic Regression was better at detecting positive cancer cases.

5.3 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.

Reference

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.

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

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.

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.

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

Kieffer, J. (1994). Elements of information theory (thomas m. Cover and joy a. Thomas). *SIAM Review*, *36*(3), 509-511.

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

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

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.

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.

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.

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

.