FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 5, May, 2025, pp 335 – 339
DOI: https://doi.org/10.33003/fjs-2025-0905-3626

CERVICAL CANCER PREDICTION USING ARTIFICIAL NEURAL NETWORKS: A CASE STUDY ON NIGERIAN HEALTHCARE DATA

***Isah, Khadijat Ozavisa, Opeyemi, Adenike Abisoye, Yemi-Peters, Victoria Ifeoluwa and Malik, Adeiza Rufai**

Department, Federal University Lokoja, Lokoja, Nigeria

*Corresponding authors' email: isah.khadijat1050@fceokene.edu.ng

ABSTRACT

Cervical cancer remains a leading cause of morbidity and mortality in low- and middle-income countries (LMICs), particularly in Nigeria, where limited healthcare resources, inadequate screening programs, and late detection exacerbate the burden. Despite advances in medical science, early detection remains a significant challenge due to socioeconomic barriers and insufficient diagnostic infrastructure. This study addresses these issues by developing a predictive model leveraging Artificial Neural Networks (ANNs) tailored to a Nigerian dataset. The model utilizes advanced data preprocessing techniques, including the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and Genetic Algorithms (GA) for feature selection, thereby optimizing the predictive accuracy and efficiency of the ANN. The dataset, consisting of 858 records of demographic, clinical, and lifestyle attributes, was preprocessed to handle missing data and normalize features. After applying SMOTE and GA, the optimized ANN model achieved an accuracy of 87%, a recall of 71%, and an Area Under the Curve (AUC) of 0.85, demonstrating significant improvements over baseline models. These results underscore the potential of machine learning to enhance cervical cancer prediction, particularly in resource-constrained settings like Nigeria. This study highlights the transformative potential of artificial intelligence (AI) in reducing healthcare disparities by providing scalable, localized solutions for early cancer detection. It also emphasizes the importance of integrating domain-specific datasets to improve model relevance and performance. The findings provide a foundation for future research aimed at deploying AI-driven healthcare systems in LMICs, with the ultimate goal of reducing cervical cancer-related mortality and improving public health outcomes.

Keywords: Artificial Neural Networks, Cervical cancer, Data preprocessing, Genetic algorithms, SMOTE, Nigerian dataset

INTRODUCTION

Cervical cancer poses a critical health challenge in low- and middle-income countries (LMICs), where limited access to screening and treatment leads to alarmingly high mortality rates. According to the World Health Organization (WHO), cervical cancer is one of the most preventable yet deadliest cancers in LMICs. It is estimated that approximately 90% of cervical cancer-related deaths occur in these regions, with sub-Saharan Africa bearing a disproportionate share of the burden (WHO, 2022). The disease continues to devastate populations in resource-limited settings, despite advancements in preventative measures like HPV vaccination and screening.

Several factors contribute to this ongoing crisis. Socioeconomic barriers, such as poverty and a lack of education, limit awareness about preventive measures and the importance of regular screening (Ronco et al., 2014). Cultural stigmas surrounding gynecological exams often discourage women from seeking care. Additionally, inadequate healthcare infrastructure further exacerbates the issue by limiting access to trained professionals, diagnostic tools, and treatment facilities (Martha et al., 20219). These challenges result in late-stage diagnoses, where treatment options are less effective, leading to high mortality rates.

In recent years, machine learning (ML) has emerged as a transformative technology in healthcare, providing scalable and cost-effective solutions for disease prediction, diagnosis, and management. ML algorithms, such as Support Vector Machines (SVM), Decision Trees, and Artificial Neural Networks (ANNs), have shown significant potential in addressing diagnostic challenges for diseases like breast cancer, cervical cancer, and other non-communicable diseases (Jiayi et al., 2023; Aljrees, 2024). Among these methods, ANNs have gained prominence for their ability to model complex, nonlinear relationships, making them particularly suitable for medical datasets with intricate interdependencies (Kaur et al., 2023). ANNs excel in processing large amounts of data, identifying hidden patterns, and providing reliable predictions that aid in clinical decision-making.

However, a significant limitation of most existing predictive models lies in their lack of generalizability to LMICs. Many of these models are developed using datasets from high-income countries, which often fail to capture the genetic, environmental, and sociocultural factors unique to populations in LMICs (Martha et al., 2019). For instance, lifestyle behaviors, access to healthcare, and genetic predispositions differ significantly between populations in high-income and low-income regions. This lack of localization in datasets hinders the applicability of machine learning models in LMICs, leaving these regions underserved despite their substantial need for reliable predictive tools.

Cervical cancer datasets present additional challenges, including class imbalance, high dimensionality, and missing data. Most medical datasets are imbalanced, with far fewer positive cases of cervical cancer compared to negative ones. This imbalance negatively impacts the model's ability to accurately predict positive cases, leading to a higher rate of false negatives, which can be life-threatening in a clinical setting. Moreover, high-dimensional datasets containing irrelevant or redundant features can reduce model efficiency and accuracy. Missing values further complicate data preprocessing, necessitating the use of robust imputation methods to maintain data quality.

In light of these challenges, this study seeks to address the gaps in cervical cancer prediction by developing an Artificial Neural Network (ANN)-based model specifically tailored to a Nigerian dataset. By leveraging advanced preprocessing techniques, including the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and Genetic Algorithms (GA) for feature optimization, the study aims to improve model performance and reliability. The dataset comprises demographic, clinical, and lifestyle attributes relevant to cervical cancer prediction in Nigeria, ensuring that the model is both localized and contextually relevant.

MATERIALS AND METHODS

This study developed a cervical cancer prediction model specifically designed for the Nigerian healthcare context using Artificial Neural Networks (ANNs). The methodology involved several key steps to ensure data quality, address class imbalance, and optimize feature selection. Fig 1 represents this study methodological framework.

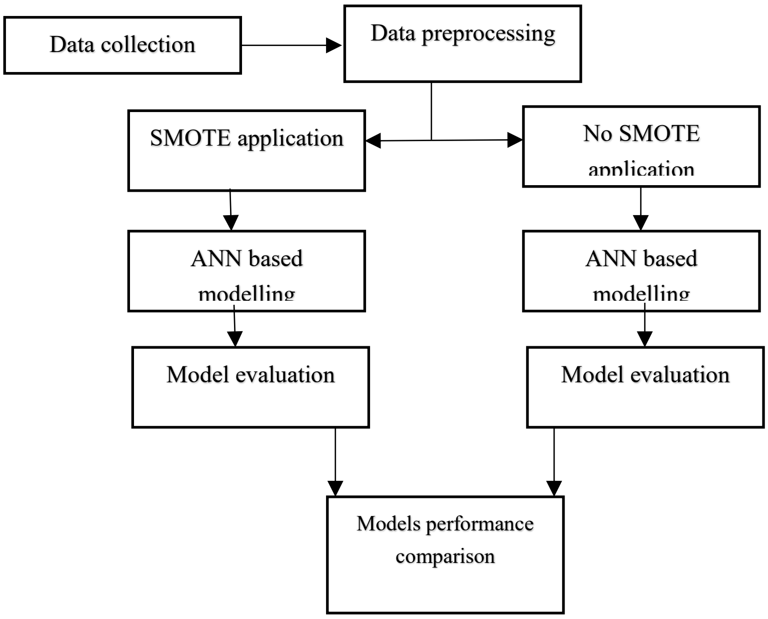

Figure 1: Study methodological framework

Data Collection

A dataset containing 858 records was collected from three (3) different hospitals in Lokoja. The dataset included demographic, clinical, and lifestyle attributes such as age, smoking history, HPV status, and sexually transmitted disease (STD) incidence. The factors considered when collecting the dataset proposed in this study were factors also considered in the dataset of Martha et al. (2019).

Data Preprocessing

To help the machine learning analytic carry out a less erroneous analysis process, the dataset collected was preprocess, where handling missing data, data normalization, handling class imbalance, data conversion (encoding) and feature selection process were carried out as the preprocessing activities.

Missing values were addressed using mean imputation through Python's SimpleImputer. Numerical features were normalized using StandardScaler to ensure uniformity. SMOTE was implemented to balance the dataset. Here. synthetic samples were generated for the minority class. Non-numeric data was converted into numerical format for ANN processing. This was done using the LabelEncoder python tool.

Genetic Algorithms (GA) were used to select the most relevant features. The algorithm refined the dataset by reducing 32 features to 15 key attributes, improving model efficiency without sacrificing accuracy in predicting cervical cancer outcomes, significantly enhancing the model's computational efficiency and predictive accuracy.

The process began by encoding the feature set as a population of candidate solutions, where each candidate represented a potential subset of features. GAs optimized this subset selection by simulating natural evolutionary processes, such as selection, crossover, and mutation. Specifically, two-point crossover and bit-flip mutation operators were applied. Two-point crossover allowed candidate solutions to exchange genetic material (i.e., feature combinations) to generate offspring solutions that combined the strengths of both parents. Bit-flip mutation introduced random changes to individual solutions, ensuring diversity and preventing premature convergence to suboptimal solutions.

To evaluate the quality of each subset, a Multi-Layer Perceptron (MLP) (ANN) classifier was trained using the selected features, and its performance was assessed through mean accuracy using 5-fold cross-validation. This approach ensured that the selected features consistently contributed to high predictive accuracy across different folds of the dataset, minimizing overfitting.

The optimization process was conducted over 10 generations, during which the GA iteratively refined feature combinations. In each generation, poorly performing subsets were discarded, while high-performing subsets were retained for further refinement. By the end of the optimization process, an optimal subset of 15 features was identified. This reduced the dimensionality of the dataset while maintaining high accuracy, recall, and F1-score.

The selected features included key predictors such as age, HPV status, smoking history, and the number of pregnancies. These variables were identified as critical based on their contribution to model performance and their known relevance in cervical cancer risk. Reducing the dimensionality of the dataset not only improved the computational efficiency of the ANN but also enhanced its interpretability, allowing healthcare professionals to better understand the model's decision-making process.

ANN based modelling

To perform an ANN based modelling on both dataset segments (preprocessed dataset with SMOTE and preprocessed dataset without SMOTE), an ANN algorithm is implemented using both dataset (as seen in fig 1). This is to measure the importance of SMOTE augmentation on the performance of an ANN model. The ANN was implemented using TensorFlow, featuring an input layer, one hidden layer with ReLU activation, and a sigmoid-activated output layer. Hyperparameters were optimized using GridSearchCV.

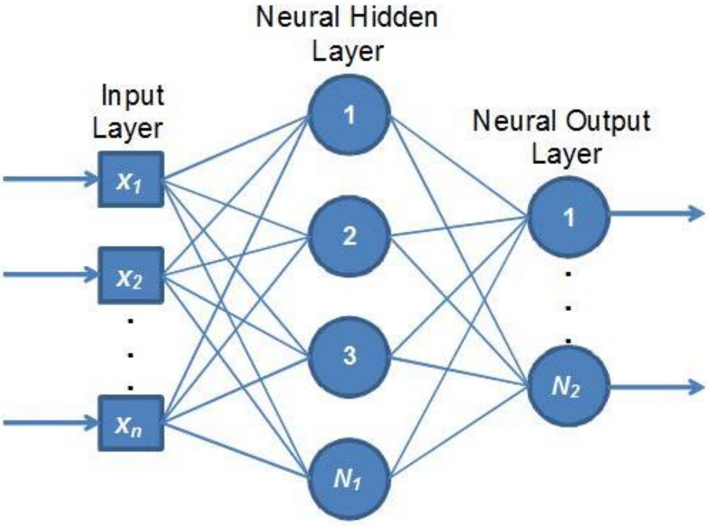

Figure 2: ANN Model Architecture (Jiayi et. al., (2020))

The ANN model was implemented using TensorFlow with the following architecture:
Input Layer: 15 neurons representing selected features.
Hidden Layer: One fully connected hidden layer with ReLU activation.
Output Layer: A single neuron with a sigmoid activation function for binary classification.
Training Algorithm: The Adam optimizer was used for training, with binary cross-entropy as the loss function.

Models Evaluation

The model’s performance was assessed using accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC). A 5-fold cross-validation approach ensured robustness and reliability.

RESULTS AND DISCUSSION

Feature Selection

The Genetic Algorithm (GA) successfully reduced the feature set from the original dataset, selecting a subset of 15 key features. This reduction in dimensionality enhanced model efficiency without compromising performance. The combination of optimized features and hyperparameter tuning further improved the ANN's reliability and generalizability. Table 1 presents the 15 key features selected using the Genetic Algorithm (GA), which enhanced the model's efficiency by reducing dimensionality without sacrificing performance.

Table 1: Key features selected using the Genetic Algorithm

Selected Features	Description
Age	Patient's age
Number of Sexual Partners	History of sexual behavior
STDs	Presence of sexually transmitted diseases
Biopsy	Target variable (positive/negative)
Hormonal Contraceptives	Use of contraceptive methods
Smoking Status	Smoking behavior
First Sexual Intercourse (Age)	Early sexual activity
IUD Usage	Use of intrauterine devices
HPV	Human Papillomavirus status
Pregnancy Count	Number of pregnancies
Marital Status	Relationship status
STD Diagnosis Count	Number of diagnosed STDs
Genital Warts	Presence of warts
STD Treatment	Received STD treatments
Education Level	Patient's education level

Model Performance

The Artificial Neural Network (ANN)-based cervical cancer prediction model was evaluated using a localized Nigerian dataset, focusing on accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). The model's performance was assessed at two stages: before and after the application of the Synthetic Minority Oversampling Technique (SMOTE) for class balancing. The performance of the Artificial Neural Network (ANN)-based cervical cancer prediction model was evaluated before and after applying the Synthetic Minority Oversampling Technique (SMOTE).

The initial evaluation revealed an accuracy of 83%, with a precision of 67% and recall of 56%. The F1-score stood at 61%, and the AUC was 0.79. These metrics indicated significant limitations in the model's ability to correctly identify the minority class (patients testing positive for cervical cancer), underscoring the need for further enhancements. Table 2 summarizes the ANN model's performance metrics before and after applying SMOTE.

Table 2: ANN performance metrics before and after SMOTE

Metric	Before SMOTE	After SMOTE
Accuracy (%)	83	87
Precision (%)	67	74
Recall (%)	56	71
F1-Score (%)	61	72
AUC	79	85

Following the application of SMOTE, the model achieved a marked improvement across all metrics. Accuracy increased to 87%, while precision and recall rose to 74% and 71%, respectively. The F1-score improved to 72%, and the AUC reached 0.85. These enhancements demonstrate that SMOTE effectively mitigated the issue of class imbalance, enabling the model to better distinguish between the majority and minority classes (Aljrees, 2024; Munshi, 2024).

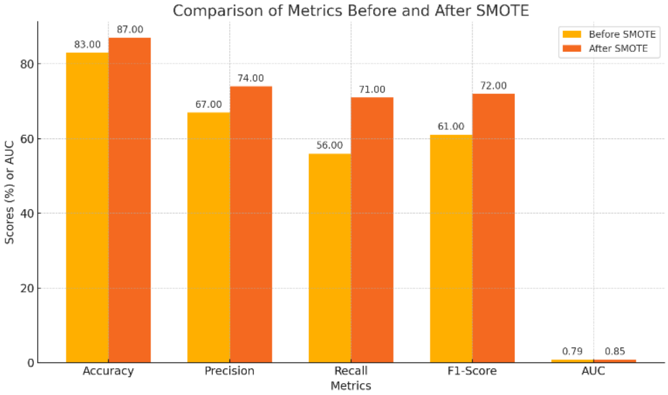

Figure 3: Performance Metrics before and after SMOTE

The improved performance metrics confirm the effectiveness of ANN in cervical cancer prediction when tailored to localize Nigerian data. The use of SMOTE effectively mitigated the problem of class imbalance, leading to better recall and reducing false negatives, which is crucial for early cancer detection, validating findings.

Feature selection using Genetic Algorithms enhanced computational efficiency by reducing unnecessary variables, ensuring a focused and accurate predictive model. This aligns with previous studies demonstrating the importance of feature selection in machine learning-based healthcare models.

Despite these improvements, the study faced limitations. The dataset size was relatively small. While SMOTE improved class balancing, it may not fully represent real-world distributions. Future research should explore the collection of much higher dataset to increase the applicability level of the model developed using the dataset, within the Nigeria context.

This study contributes to the growing body of knowledge on AI-driven healthcare solutions in LMICs. The findings affirm that localized, data-driven models can significantly enhance early detection of cervical cancer, potentially reducing its burden in resource-constrained settings. Future research should focus on integrating clinical trials and real-time data to validate the model's effectiveness in practical healthcare applications.

CONCLUSION

This study developed an ANN-based cervical cancer prediction model tailored to Nigerian healthcare data. The use of SMOTE improved the model's ability to identify positive cases, while Genetic Algorithms optimized feature selection, ensuring efficiency without sacrificing accuracy. The findings highlight the potential of machine learning in enhancing early cervical cancer detection, particularly in resource-limited settings. The integration of SMOTE addressed class imbalance, and GA-driven feature selection reduced computational complexity while maintaining performance.

FURTHER WORK

While this study provides a foundational approach to cervical cancer prediction in Nigeria, several areas warrant further exploration to enhance the model's generalizability, applicability, and real-world impact:

i. Expanding the Dataset: This study relied on a limited dataset collected from online sources. Expanding the dataset to include data from multiple Nigerian regions, as well as incorporating more diverse demographic and lifestyle factors, would enhance the robustness and representativeness of the model.

ii. Real-World Implementation: Implementing the model in real-world healthcare settings, such as local hospitals and clinics, would provide valuable insights into its usability, reliability, and scalability. Feedback from healthcare professionals and patients would be essential for fine-tuning the system to meet the practical needs of the medical community.

iii. Interpretability and Explainability: A key area for future work is the interpretability and explainability of the ANN model, especially in medical applications. Utilizing explainable AI (XAI) methods would enhance trust and transparency in the predictions, allowing healthcare professionals to better understand how the model makes its decisions. This is crucial for improving model adoption in clinical environments.

In conclusion, this study lays the groundwork for leveraging machine learning to improve cervical cancer detection in Nigeria and similar low-resource settings. The promising results highlight the transformative potential of AI in public health, but continued research and collaboration with healthcare professionals are essential to further optimize and implement such technologies in real-world applications.

REFERENCES

Alassaf, A., Alarbeed, E., Alrasheed, G., Almirdasie, A., Almutairi, S., Al-Hagery, M. A., & Saeed, F. (2024). Genetic algorithms and feature selection for improving the classification performance in healthcare. *International Journal of Advanced Computer Science and Applications, 15*(3), 737-744. https://doi.org/10.14569/IJACSA.2024.0150375

Aljrees, T. (2024). Cervical cancer detection using K-nearest neighbor imputer and stacked ensemble learning model. DIGITAL HEALTH, 9. https://doi.org/10.1177/20552076231203802

Bae, S., et al. (2020). The role of deep learning in early cancer detection. IEEE Access, 8, 16850-16857.

Chanudom, I., Tharavichitkul, E., & Laosiritaworn, W. (2024). Prediction of cervical cancer patients' survival period with machine learning techniques. Healthcare Informatics Research, 30(1), 60-72. https://doi.org/10.4258/hir.2024.30.1.60

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. https://doi.org/10.1613/jair.953

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2), 182-197. https://doi.org/10.1109/4235.996017

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115–118. https://doi.org/10.1038/nature21056

Ghaheri, A., Shoar, S., Naderan, M., & Hoseini, S. S. (2015). The applications of genetic algorithms in medicine. Oman Medical Journal, 30(6), 406-416. https://doi.org/10.5001/omj.2015.82

He, H., & Garcia, E. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Jiayi, L., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Generation Computer Systems, 106, 199–205. https://doi.org/10.1016/j.future.2019.12.033

Kaur, S., Sharma, L. M., Mishra, V., Goyal, M. G. B., Swasti, S., Talele, A., & Parikh, P. M. (2023). Challenges in cervical cancer prevention: Real-world scenario in India. South Asian Journal of Cancer, 12(1), 9-16. https://doi.org/10.1055/s-0043-1764222

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005

Martha, W. M., Kyama, M. C., & Kibert, P. S. (2019). Improving early diagnosis of cervical cancer lesions using p16INK4a biomarkers on cell blocks from cervical smears. African Journal of Health Sciences, 32(2).

Munshi, R. M. (2024). Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. PLOS ONE, 19(1), e0296107. https://doi.org/10.1371/journal.pone.0296107

Ramos-Pollán, R., de la Vega, R., Álvarez-Fernández, A., López-Morales, M., & Cossío, J. A. (2019). Breast cancer classification using machine learning. PLOS ONE, 14(9), e0221581. https://doi.org/10.1371/journal.pone.0221581

Ronco, G., Giorgi-Rossi, P., Carozzi, F., & Dunne, M. (2014). Conventional screening in low-resource settings. Cervical Cancer Screening: Past, Present, and Future, 5(2), 75-83. https://doi.org/10.3802/jgo.2014.5.2.75

Rupali, V., Handa, R., & Puri, V. (2020). Feature selection using genetic algorithm for cancer prediction system. In Proceedings of the International Conference on Computational Science and Engineering. https://doi.org/10.1007/978-981-15-5341-7_91

World Health Organization. (2022). Cervical cancer statistics and global burden. Retrieved from https://www.who.int/publications/i/item/cervical-cancer-statistics-and-global-burden