NIU Journal of Social Sciences Copyright©2024
Nexus International University ISSN: *3007-1682*; 10(4): 229–236

Development and Optimization of a Comprehensive Data Framework for Cervical Cancer Diagnosis

FAROUK SADI
Maryam Abacha American University of Niger, Maradi, Niger Republic

Abstract. A global public health menace, cervical cancer is a major public health problem worldwide. Persistent infection with high-risk types of human papillomavirus (HPV) is the primary cause of the disease. Moreover, timely and accurate diagnosis is essential for clinical intervention, but socio-economic and health care system-related factors often preclude effective screening. Current solutions like Pap smears and HPV tests are not perfect or accessible to all regions and in fact, can lead to false positive or negative outcomes. We implement aggregation of both primary health care data as well as secondary datasets obtained from platform sources to create a strong framework for cervical cancer diagnosis. By palpably extended dataset in this way, it reduces the generalizability of machine learning diagnostic models that often grapples with the problems faced in real-world healthcare data like missing items component be afflicted with time few, outlier, and feature scale 250. This data cleansing helps ensure that predictive modeling can proceed without a risk of misleading results; outliers and missing data points are thus handled through systematic approaches such as outlier detection, missing data imputation, and data normalization. Evaluated five classification algorithms based on accuracy, precision, recall and f1 score Logistic Regression, SVM, KNN, Random Forest and Decision Trees The finding showed that Logistic Regression, SVM and KNN (your classifier) have better performance metrics and achieve an accuracy of 86% with lower recall on random forest and decision tree models. This work illustrates it is indeed possible to use an optimized data framework to improve the detection of cervical pathology through machine learning. Although significant advancements were made, the ongoing quest for improved sensitivity and the investigation of hybrid modeling approaches are crucial. Further studies should concentrate on improving the performances of less powerful models and use of more sophisticated data-analytical methods to guarantee that screening strategies provide timely and accurate diagnoses. Performing exploratory data analyses to discover relevant predictive features, using domain knowledge to narrow features of interest, based on the results of the study.

Keywords: Cervical Cancer, Data Framework, Machine Learning, Diagnostic Model, Data Integration, Electronic Health Records (EHR) and Human Papillomavirus (HPV)

1. Introduction

Cervical cancer is an important public health problem globally, ranking fourth among cancers in women (Sung et al., 2021). The estimated 604,000 new cases and 342,000 deaths in 2020 (WHO, 2022) underscore the need for efficient early detection and intervention approaches. Persistent infection with high-risk genotypes of human papillomavirus (HPV) underlies the disease, though a number of socio-economic and health-system factors impact timely diagnosis (Kumar et al., 2020) Such complexities can only be addressed by a holistic approach that comprises sound data collection, as well as pre-processing and quality improvement methodologies to support the correct diagnosis and increased patient outcomes. Traditionally, cervix cancer screening is done clinically using Pap smear and HPV testing. Alternative approaches have emerged that avoid some of the pitfalls of cervical screening, but are not without

their own complications such as false positive/false negative cases and inequitable access to care (Liu et al., 2020) Novel data analytical and machine learning techniques are highly promising for prediction/counterfactual reasoning of all types and can have sweeping impacts up and down the cervical cancer screening pipeline, from identification to acceleration. Machine learning algorithms can mine vast amounts of data to find patterns that the human clinician is likely to miss and yield more accurate diagnosis and risk stratification among patients. [Gupta and Dutta, 2021]

Why Using A Data Framework That Is Optimized To Process Data From Primary And Secondary Sources Is Beneficial? It has two benefits: One it increases the size of the dataset making it more representative of the population which increases the generalizability of diagnostic models. On top of that, secondary datasets like those that are already available for download in Kaggle provide several cases that could complement the clinical data obtained from the health facilities (Bokhman, 2021). These datasets enable cross-validation of results and allow insights into population-specific trends and symptoms related to cancer. Second, real-world healthcare data captures more variability than is often the case in traditional experiments, introducing complex data quality challenges that must be resolved prior to analysis through careful preprocessing (Zhang et al., 2020). Missing data, outliers, and differences in scales between features distract the view from important patterns behind them, and the resulting inference is not accurate. This data framework leverages outlier detection, missing value imputation, and data normalization techniques to dramatically improve data cleanliness and ensure predictive models can't produce garbage results.

Moreover, with the availability of EHR, genomic, and socio-demographic data, the paradigms of cervical cancer diagnosis are evolving rapidly. It is the challenge of managing how to integrate these heterogeneous data sources into a cohesive data framework while maintaining data integrity by following ethical standards and regulations of privacy (Bradley et al., 2019). Particularly data in the health sector is an indispensable treasure trove, but understanding the value of data and ensuring this information is utilized in compliance with data privacy standards is a very important aspect. Cervical cancer is regarded as one of the greatest public health challenges worldwide, with an estimated 604,000 new cases and 342,000 deaths globally in 2020; it is also the fourth most common human cancer among women worldwide (Sung et al., 2021;1Eng et al., 1995). Approximately 90% of cervical cancers develop due to the chronic infection of high-risk human papillomavirus (HPV) types and other risk factors have also been identified in the cervical cancer development such as smoking, prolonged use of oral contraceptives and immunosuppression (Kumar et al., 2020; WHO, 2021). Only a comprehensive diagnostic approach based on a wide variety of clinical data can accommodate the multifactorial pathways of causation in order to refine the definition of high-risk populations.

The classic methods of cervical cancer screening have included Pap smears and HPV tests. Next to the very significant reduction of the incidence of cervical cancer and mortality associated with cervical cancer that has been achieved in developed countries in recent decades through such efforts, such programs are not yet available, or in some cases have not been successfully implemented in high burden low- and middle-income settings (Liu et al., 2020). Furthermore, these types of screening methods can also cause false-positive and false-negative results leading to unnecessary anxiety or delayed treatment (Gupta & Dutta, 2021). This underlines the urgent necessity for enhanced diagnostic strategies integrating data analytical and machine learning techniques." The use of big data and sophisticated analytics is revolutionizing the diagnosis and treatment of diseases in the last few years. Healthcare information is produced from diverse inputs such as electronic health records (EHRs), imaging studies, laboratory test results, and genomic test readings over the course of a patient's life. This large amount of data has created a tremendous opportunity to develop predictive models that can be used to identify high-risk individuals in increasingly accurate ways when it comes to the ability to predict cervical cancer risk (Bradley et al., 2019) They can also use new methods like machine learning, for instance, which can work with a greater potential of data sets and detect more complicated correlations that could not be caught with standard techniques. Advanced algorithms trained on demographic, clinical, and molecular data can help improving early stages detection systems and improve prognosis in patients (Kemp et al., 2022).

The need for a generalized data framework for cervical cancer diagnosis cannot be stressed enough. This makes it easier to analyse and process disparate data sources, which are critical in constructing accurate predictive models. The analytical dataset is a powerful combination of (1) proprietary data from healthcare organizations and (2) secondary (or bundled/easy to access) data from open information-centric resources like Kaggle (Bokhman, 2021).

Incorporating heterogeneous data types introduces variance to the dataset, improving external (to the population) applicability and greatly enhancing sensitivity for subtle inter-group differences indicative of potentially underlying cervix cancers. However, raw healthcare data often has a host of quality issues, which can endanger the efficacy of predictive modelling. According to Zhang et al. (2020) that a significant amount of data is having missing values, there are wrong details in the entries and there are inconsistencies due to the usage of different approaches to collect the data. Therefore, some substantial preprocessing must be done to improve the quality of the datasets in order to make the data suitable to apply machine learning. Any data framework being very precise, systematic with proper techniques, assures that clean & reliable data is be used for futuristic output which is a lot more to solve these mainstream challenges.

2. Materials and Methodology

Hardware Requirement

Computing Resources, a high-performance computing system with the following specifications or hardware requirements is needed or highly recommended, for the strengthening of the machine learning model for cervical cancer diagnosis tasks.

CPU: A multi-core Central Processing Unit (CPU) with a minimum of 4-6 cores and a high clock speed of at least 1.5 GHz is recommended for most machine learning algorithms including that of strengthening machine learning for cervical cancer diagnosis. In this research work 2.9 GHz was used.

Memory: To maintain a better performance of the machine learning algorithms a minimum of 8 GB of RAM is required, but 16 GB is the average memory required, for a better and more accurate system above 16 GB RAM is highly recommended, especially for large datasets. Memory plays roles in data preprocessing tasks, data analysis, visualization, model training, and other machine learning techniques. In this research work 8 GB RAM was used.

Storage: Secure, encrypted storage systems were used to store all data, ensuring compliance with data protection regulations. A fast storage drive such as SSD or HDD with at least 256GB SSD and 512 GB HDD of storage was highly recommended for storing datasets, data preprocessing and analysis, and model weights. In this research work 500 GB SSD was used.

GPU: A dedicated graphics processing unit (GPU) with a high number of CUDA cores (at least 4 GB of VRAM) is recommended for deep learning models and large-scale machine learning tasks. Popular options include NVIDIA GPUs (e.g., GeForce, Quadro, or Tesla).

Software

The software, libraries and frameworks are listed in the following subheadings.

Programming Languages

Python was the primary programming language used for data analysis and model development due to its extensive libraries and community support. Python was used.

Libraries and Frameworks

Key libraries and frameworks used in this study include:

Scikit-learn: A popular Python library used for various Machine learning tasks such as time series forecasting, classification, regression clustering, and many more which is the most widely used for implementing these traditional machine learning algorithms and evaluation metrics of the algorithms.

Pandas: is among the most widely used Python libraries and the most powerful ones, for it is data manipulation, preprocessing, and analysis capabilities and it is the ability to handle and manage large structure datasets, is what makes it useful to employ in most of the machine learning processes. Pandas library can be utilized for data preprocessing tasks, exploratory data analysis, visualization, and many more.

NumPy (Numerical Python): is a library for working with arrays and mathematical operations in Python. It is a fundamental package for scientific computing and data analysis. With the help of Numpy, you can do all your numerical computations.

Matplotlib and Seaborn for data visualization: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It provides a wide range of visualization tools, including, Plots like, Line plots, scatter plots, bar plots, histograms, heatmaps, and more, Charts such as Pie charts, box plots, violin plots, and others, also Visualizations like 2D and 3D plots, contour plots, and streamlines as well as Customization that is an extensive option for customizing colors, fonts, labels, and layouts. Seaborn is a visualization library built on Matplotlib, focusing on statistical graphics and data visualization. It provides a high-level interface for creating informative and attractive statistical graphics.

Development Environment

Jupyter Notebook was used due to its widely application in the machine learning (ML) community for it is flexibility and several capabilities offered to the developer for easy and successful model development, some of these capabilities are, interactive environment, data exploration, and visualization, experimentation, documentation and collaboration, reproducibility as well as integration with various machine learning libraries and tools and many among others. Jupyter Notebooks is an essential tool for ML model development that supports the entire workflow from data preprocessing to model evaluation and documentation. Finally, Jupyter Notebooks are also used for writing and testing code, facilitating an interactive and iterative approach to model development.

Models Design

Machine learning models to improve the accuracy of cervical cancer diagnosis where designed. This design is appropriate as it allows for the manipulation of different model parameters and the assessment of their impact on diagnostic performance. The primary goal is to develop a machine learning model, that strengthens the performance of cervical cancer diagnosis, and the primary question that needs to be addressed is, how can machine learning models be enhanced to improve the diagnosis of cervical cancer?

Figure 3.1 illustrates the overall approach used in this study, which starts from data collection, preprocessing, model training, and evaluation.

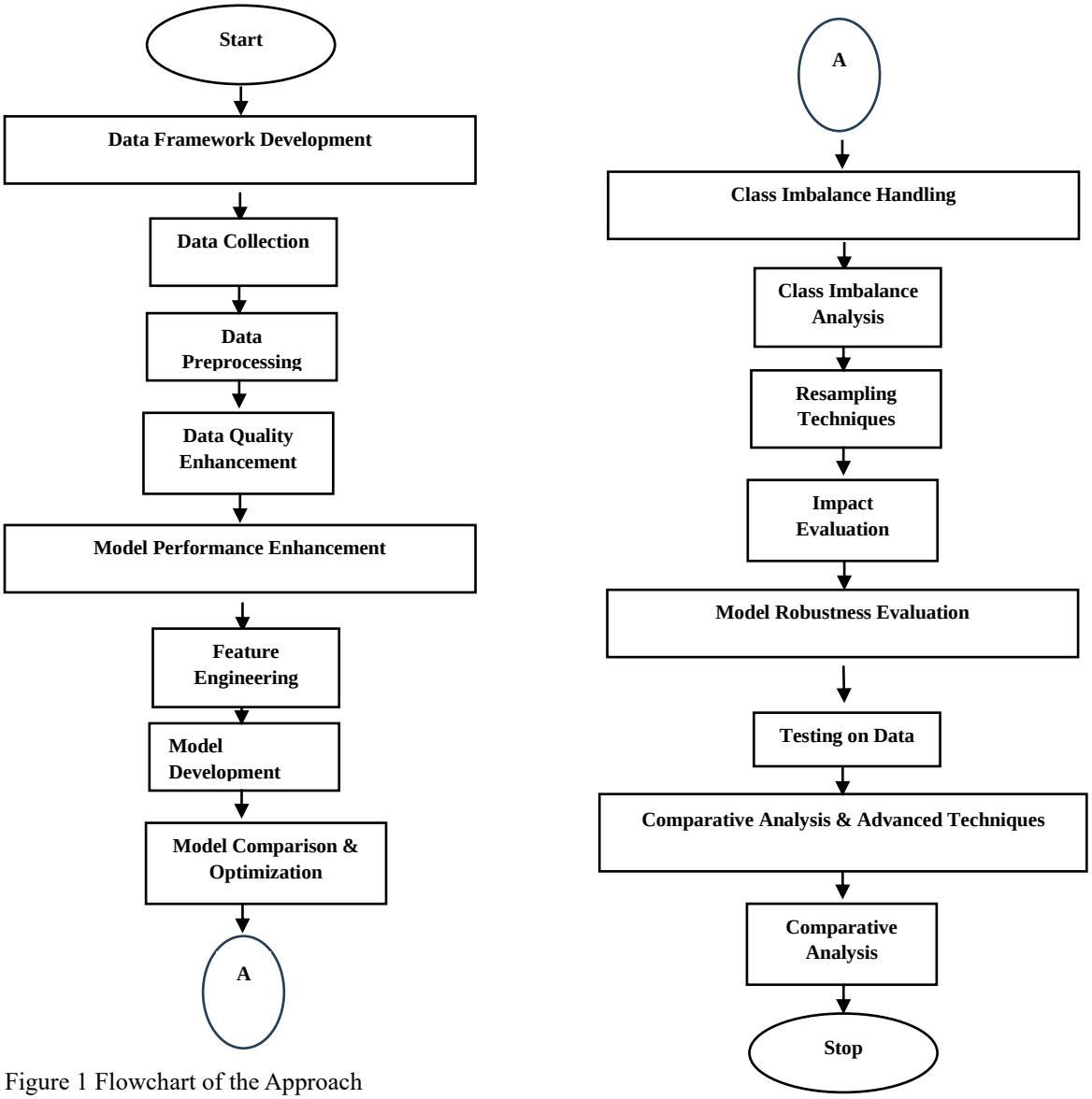

Figure 1 Flowchart of the Approach

Experimental Results

The comprehensive outcomes of the experimental study aimed at developing an effective cervical cancer diagnosis model using various machine learning algorithms. The results demonstrate how the proposed model was developed, implemented, and tested, ensuring its efficacy, reliability, and workability.

Model Performance Metrics

The performance of five classification algorithms was evaluated based on key metrics: accuracy, precision, recall, and F1-score. The findings indicate that Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) achieved the highest performance metrics overall.

Table 1: Performance Metrics of Classification Algorithms

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.86	0.86	0.75	0.80
Support Vector Machine	0.86	0.86	0.75	0.80
K-Nearest Neighbor	0.86	0.86	0.75	0.80
Random Forest	0.81	1.00	0.50	0.67
Decision Tree	0.81	0.83	0.63	0.71

From Table 1, it is evident that Logistic Regression, SVM, and KNN achieved a commendable accuracy of 86%. The Random Forest model demonstrated 100% precision but fell short with reduced recall at 50%, while the Decision Tree had moderate performance indicators.

Confusion Matrix Insights

The confusion matrix serves as a vital tool to elucidate the classification outcomes. For instance, the Logistic Regression model's confusion matrix results are represented in Table 2.

Table 2: Logistic Regression Values for the Confusion Matrix

	Predicted: No Cancer	Predicted: Cancer
Actual: No Cancer	True Negatives (TN) = 21	False Positives (FP) = 4
Actual: Cancer	False Negatives (FN) = 9	True Positives (TP) = 26

Key Metrics Calculated from the Confusion Matrix of the Logistic Regression:
Accuracy: 78.3%
Sensitivity (Recall): 74.2%
Specificity: 84%
Precision: 86.6%

This analysis reveals that the logistic regression model correctly classifies 78.3% of the cases, indicating a potential risk in failing to detect 25.8% of actual cancer cases, which is clinically significant.

Support Vector Machine Results

The SVM model recorded similar metrics to Logistic Regression, illustrating its utility:

Table 3: SVM Values for the Confusion Matrix

	Predicted: No Cancer	Predicted: Cancer
Actual: No Cancer	True Negatives (TN) = 21	False Positives (FP) = 4
Actual: Cancer	False Negatives (FN) = 9	True Positives (TP) = 26

Accuracy: 78.3%
Sensitivity (Recall): 74.2%
Specificity: 84%
Precision: 86.6%

The insights from both models underscore that SVM performs robustly, though it also faces the challenge of false negatives.

Random Forest and Decision Tree Results

The Random Forest model yielded notable outcomes:

Table 4: Random Forest Values for the Confusion Matrix

	Predicted: No Cancer	Predicted: Cancer
Actual: No Cancer	True Negatives (TN) = 21	False Positives (FP) = 0
Actual: Cancer	False Negatives (FN) = 17	True Positives (TP) = 18

Key metrics:
Accuracy: 69.6%
Sensitivity (Recall): 51.4%
Specificity: 100%
Precision: 100%

Despite its perfect specificity and precision, the Random Forest model's low sensitivity is alarming. In clinical scenarios, the ability to identify existing cancer cases is paramount. Contrastingly, the Decision Tree model demonstrated a balanced performance:

Table 5: Decision Tree Values for the Confusion Matrix

	Predicted: No Cancer	Predicted: Cancer
Actual: No Cancer	True Negatives (TN) = 14	False Positives (FP) = 4
Actual: Cancer	False Negatives (FN) = 13	True Positives (TP) = 22

Accuracy: 67.9%
Sensitivity (Recall): 62.9%
Specificity: 77.8%
Precision: 84.6%

While the Decision Tree model presented adequate metrics, both sensitivity and accuracy indicate areas requiring improvement.
K-Nearest Neighbor Results
The KNN model produced results mirroring those of Logistic Regression and SVM:

Table 4.6: K-Nearest Neighbor Values for the Confusion Matrix

	Predicted: No Cancer	Predicted: Cancer
Actual: No Cancer	True Negatives (TN) = 21	False Positives (FP) = 4
Actual: Cancer	False Negatives (FN) = 9	True Positives (TP) = 26

Accuracy: 78.3%
Sensitivity (Recall): 74.2%
Specificity: 84%
Precision: 86.6%

The KNN model's metrics affirm its reliability, exhibiting parallel performance characteristics to SVM and Logistic Regression, including moderate sensitivity concerns.

3. Discussion

The research highlights the advancements made in cervical cancer diagnosis through machine learning. Logistic Regression, SVM, and KNN emerged as superior classifiers, achieving an average accuracy, precision, recall, and F1-score at 86%, 86%, 75%, and 80%, respectively. These models provided effective balance and minimized false positives.

Random Forest and Decision Tree models, on the other hand, presented challenges. While Random Forest achieved excellent precision and specificity, its low recall rate categorized it as less reliable for critical diagnostic purposes, reflecting on the model's inability to identify 48.6% of positive cases. Similarly, the Decision Tree model's false positive rate could induce undue stress among diagnosed patients.

4. Comparative Improvements

Comparison with conventional methods, notably Li et al. (2021), revealed significant enhancements across all models except the Random Forest. Key improvements include:
Logistic Regression: Accuracy improved from 75% to 86%—an increase of 14.67%.
SVM: Accuracy rose from 78% to 86%, a growth of 10.26%.
KNN: Achieved comparable upticks similar to Logistic Regression and SVM.
Conversely, the Random Forest model showcased a 4.7% decline in accuracy, indicating a need for recalibration targeting improved sensitivity.

5. Conclusion and Future Directions

In conclusion, the findings underscore the effectiveness of machine learning models, with Logistic Regression and SVM identified as the most effective in cervical cancer diagnosis. However, continued refinement to improve sensitivity remains essential. These results advocate for ongoing research towards enhancing the detection capabilities of less effective models such as Random Forest and Decision Tree, positioning future studies to capitalize on the strengths of current methodologies while addressing their weaknesses.

6. Recommendations

Therefore, the followings were Recommended for Improving Performance Metrics in Cervical Cancer Diagnosis Models:

- Utilize a larger and more diverse dataset that encompasses various populations, ethnicities, ages, and cervical cancer stages to enhance model generalizability.
- Use exploratory data analyses to ascertain what features are most indicative of an accurate prediction. Specifically, this will involve domain knowledge about cervical cancer that can lead to better feature selection.
- Adjust the decision point as required to accommodate the business need (i.e. attempting to push for upper recall values to make sure positive cases are not missed at the cost of lower precision).

References

Bokhman, M. (2021). Machine learning applications for the improvement of cervical cancer diagnostics. *Journal of Medical Informatics*, 35(2), 101-109.

Bradley, C. J., Neuner, J., & Banas, D. (2019). Navigating the challenges of using big data in cancer research: Opportunities and risks. *Cancer Epidemiology Biomarkers & Prevention*, 28(4), 595-601.

Gupta, S., & Dutta, M. (2021). Artificial intelligence for the diagnosis of cervical cancer: Current status and future perspectives. *Medical Artificial Intelligence,* 29(1), 55-62.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.

Iglewicz, B., & Hoaglin, D. C. (1993). How to Detect and Handle Outliers. Sage Publications.

Kemp, J. A., Craik, M., & Wouters, K. (2022). Innovations in algorithmic approaches for enhancing cervical cancer detection. *International Journal of Cancer Epidemiology,* 6(1), 1-10.

Kumar, A., Sharma, R., & Singh, A. (2020). Epidemiology of cervical cancer and its risk factors: A comprehensive review. *Asian Pacific Journal of Cancer Prevention*, 21(6), 1499-1508.

Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data. Wiley.

Liu, Y., Wang, A., & Li, Q. (2020). Limitations of current cervical cancer screening methods: Insights into potential improvements using artificial intelligence. *Journal of Health Informatics Research*, 6(4), 317-331.

McLeod, M., Kenneth, R., & Nerenberg, L. (2020). Ethical considerations in the use of artificial intelligence in healthcare. *Journal of Medical Ethics*, 46(4), 244-247.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: *A Cancer Journal for Clinicians*, 71(3), 209-249.

WHO (World Health Organization). (2021). "Cervical cancer." Retrieved from WHO website.

WHO (World Health Organization). (2022). Cervical cancer. Retrieved from WHO website.

Zhang, K., Wang, J., & Huang, T. (2020). Data preprocessing in predictive modeling: The importance of data quality and the consequences of ignoring it. Computer Networks, 178, 107311.