Comparative Analysis of Machine Learning Algorithms for Cervical Cancer Prediction

Ireddi Rakshitha[1], Thokala Vasisri[1], Mayaluri Anusha[1], M. Sucharitha[1*]

[1]VIT-AP University, Amaravati, Andhra Pradesh, India

sucharitha.jackson@vitap.ac.in

Abstract. Cervical cancer constitutes a significant public health concern and early diagnosis plays an important role in the patient's recovery. In this study, we investigated the utilization of various algorithms in machine learning to predict cancer with best accuracy. The objective of the paper is to identify the most reliable predictors of cervical cancer through comparative analysis. To achieve this goal, we obtained information including medical and demographic characteristics of different patients. The data has been prepared for analysis by addressing any missing values, normalizing features, and by resolving intra-class imbalance. We used algorithms like Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Decision Tree, Random Forest, XG Boost etc. Metrics like precision, accuracy, and recall, and area under receiver operating characteristic curve (AUC-ROC) are used for evaluating accuracy and discrimination. Performance of these models is also compared to real- world applications. We highlight significance of machine learning algorithms in early prediction of cervical cancer. Among all the models used, XG Boost is getting higher accuracy of 99.22%. These findings provide valuable insights to researchers, physicians and policy makers, leading to ways to enhance care for patient and to mitigate the global impact of cervical cancer worldwide.

Keywords: Performance, Algorithms, Predictive Modeling

1 Introduction

This paper offers an extensive examination of machine learning algorithms for predicting cervical cancer. Cervical cancer has the second-highest mortality rate among women in developing countries, following breast cancer [1]. It remains a global health problem, emphasizing the need for accurate and timely screening. Machine learning techniques have displayed potential in many fields of medicine, providing the ability to improve predictive models. The aim of this paper is comparing and evaluating effectiveness of different systems in cancer prediction. Our aim is to identify methods that provide the highest accuracy and reliable estimates by analyzing different data including clinical and demographic characteristics. In pursuit of this objective, we meticulously preprocess the data to address any missing values, normalize the dataset, and manage class imbalances. Then, popular machine learning algorithms like logistic

C. Kiran Mai et al. (eds.), *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*, Atlantis Highlights in Computer Sciences 18,
https://doi.org/10.2991/978-94-6463-314-6_16

regression, support vector machines, random forests, gradient boosting, and neural networks are used to build predictive models. Performance evaluation using cross-validation methods, considering criteria of key evaluation metrics which encompass accuracy, precision, recall, and the area under the receiver operating characteristics curve (AUC-ROC). In addition, computational efficiency is taken into account to examine the utility of the model in a genuine clinical setting.

The outcomes of this research will offer deeper insights into the efficacy of various learning systems in the prediction of cervical cancer. Analytical techniques can improve the accuracy of diagnosis, facilitate early diagnosis, and suggest timely interventions.

2 Literature Review

Cervical cancer stands as a prominent and pervasive health issue, particularly in economically challenged regions such as India, where it continues to claim the lives of women prematurely, making it a leading cause of female mortality globally [2]. Several research papers have explored various aspects of cervical cancer, including its etiology, risk factors, screening methods, diagnosis, treatment options, and prevention strategies [3]. Following an in-depth investigation, certain researchers have drawn the conclusion that machine learning holds boundless potential within the realm of medical research [4].

2.1 Etiology and Risk Factors

Multiple investigations have confirmed HPV infection as the foremost causative element for cervical cancer [5]. Research has focused on identifying different HPV genotypes associated with cervical cancer and their prevalence. Additionally, investigations into other risk factors such as smoking, hormonal contraceptives, immunodeficiency, and socioeconomic factors have helped identify high-risk populations and develop targeted prevention strategies [6].

2.2 Screening and Early Detection

It typically requires a span of 10 to 15 years for cervical cancer to develop, hence the timely prognosis and diagnosis of cervical cancer can save lives [7]. Cervical cancer screening initiatives utilizing approaches such as Pap smears, HPV DNA testing, and liquid-based cytology have notably decreased the incidence and mortality rates associated with cervical cancer [8]. Research has evaluated the accuracy, cost-effectiveness, and implementation strategies of these screening methods to enhance early detection rates [9].

2.3 Diagnostic Techniques

Advancements in diagnostic techniques have improved early and accurate diagnosis of cervical cancer [10]. Studies have examined the role of colposcopy, biopsy, histopathology, and molecular biomarkers in enhancing diagnostic accuracy [11]. Non-invasive methods like liquid biopsy and novel imaging modalities have also shown promise for early- stage diagnosis and treatment monitoring.

2.4 Treatment Strategies

Numerous research articles have comprehensively addressed diverse treatment modalities for cervical cancer, which encompass surgical interventions, radiation therapy, chemotherapy, and targeted therapies. Comparative studies have evaluated the efficacy, safety, and long-term outcomes of different treatment modalities, aiding in personalized treatment plans [12].

2.5 Prevention and Vaccination

Immunization against high-risk HPV types has demonstrated significant progress in cervical cancer prevention. Research findings have established the efficacy of HPV vaccines in diminishing rates of HPV infection, precancerous lesions, and the incidence of cervical cancer. Additionally, studies have focused on vaccination strategies, coverage rates, and the impact of vaccination on herd immunity [13].

2.6 Health Education and Behavioral Interventions

Understanding the social and behavioral factors influencing cervical cancer prevention and screening is crucial. Research has investigated the knowledge, attitudes, and practices related to cervical cancer among various populations, leading to the development of targeted health education programs and behavioral interventions [14].

Existing survey have provided valuable insights into the etiology, risk factors, screening methods, diagnostic techniques, treatment strategies, and prevention approaches for cervical cancer. According to a paper published recently in 2023, Machine learning techniques have demonstrated their efficacy in handling the intricacies of extensive datasets and identifying predictive attributes [15]. The results have provided the foundation for the advancement of more effective strategies to combat cervical cancer globally, ultimately aiming to reduce its incidence and mortality rates.

3 Methodology

3.1 Dataset

The dataset titled "kag_risk_factors_cervical_cancer" contains valuable information pertaining to cervical cancer risk and medical history. It includes features like age, the count of sexual partners, and the age at initial sexual activity, number of pregnancies,

smoking status, hormonal contraceptive use, intrauterine device use, history of sexually transmitted disease (STD) and various other metrics. This information provides information about uterine cancer and risk factors for its development. We use this information to apply machine learning techniques to classify cancer cells. We have taken steps before processing the data to ensure its quality and reliability.

3.2 Pre-Processing

In the given cervical cancer dataset, we took the necessary steps to handle missing values by removing them and replacing them with appropriate mean values. Missing values can significantly affect the analysis and modeling process, leading to biased results or incomplete insights. Therefore, we employed a systematic approach to deal with missing data. First, we identified the variables that contained missing values and assessed the extent of missing across the dataset. Next, we decided to remove any unnecessary columns that did not contribute significantly to our analysis or modeling goals. This process facilitated dataset refinement, concentrating on the most pertinent features. Any remaining variables with missing values were imputed with the mean value of their respective features. By using mean as a replacement, we preserved the overall distribution and statistical properties of the data. This approach allowed us to ensure data completeness and maintain the integrity of the dataset while minimizing the impact of missing values on subsequent analyses or machine learning models.

Feature Scaling. In order to prepare the cervical cancer dataset for modeling, we performed scaling of features. We employed common feature scaling methods such as standardization (z-score normalization) or normalization (min-max scaling) to transform the features into a consistent scale, allowing us to make more meaningful comparisons and interpretations in our analysis. By performing feature scaling on the cervical cancer dataset, we optimized the data for effective modeling and achieved better results in our subsequent machine learning tasks.

Feature Selection. After performing feature scaling on the cervical cancer dataset, we proceeded with selection of features. In context of cervical cancer dataset, selection of features allowed us for identifying the subset of features that had the most significant impact on predicting the presence or absence of cervical cancer. In our analysis, we utilized several feature selection methods, including correlation analysis, mutual information, and statistical tests, to assess the significance of each feature. Additionally, we considered domain knowledge and expert insights to guide our feature selection process. By carefully selecting features, we built an efficient model.

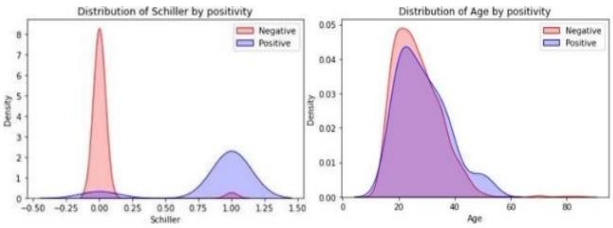

Fig. 1. Distribution of Schiller variable and Age Variable respectively

The above figure 1 shows the KDE plot to visualize the distribution of the "Schiller" variable and “Age” variable in a data frame, split by positive and negative cases.

3.3 Choosing Predictive model for Analysis

Considering the diversity of algorithms and methods, choosing the appropriate measurement model is an important step in our article. It is important to carefully evaluate and select models that suit our specific goals and profile characteristics. To ensure the best choice, we first clearly understand the problem at hand and narrow down the range of models that can achieve our goals. To determine which model is best, we consider profile characteristics such as profile type, distribution, and relationship between variables. We then evaluate the suitability of different algorithms, including terms such as interpretability, efficiency, scalability, and handling of high or unstable data. To make informed decisions, we tested various models and compared them using appropriate metrics. We maintain a balance between interpretation and reality by considering the complexity of the model, the continuous process of operation, and the incorporation of domain knowledge. We also use methods such as competition to evaluate the model's ability and reduce its workload. Following these principles allows us to choose the most appropriate evaluation model for our articles and get the best possible results.

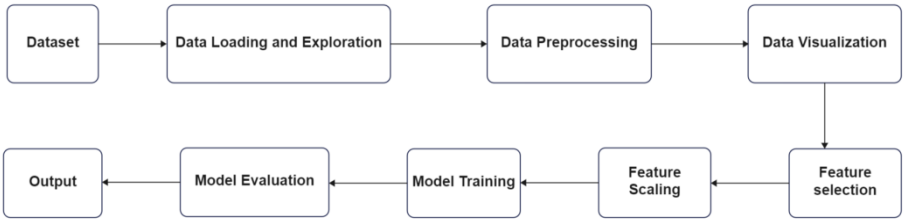

Fig. 2. Proposed Method

This flowchart outlines a proposed framework for predicting cervical cancer risk. Starting with data loading and exploration, it ensures an understanding of the dataset's structure. Preprocessing involves handling missing values and converting data types. Data visualization provides insights into age distribution and cancer prevalence.

Feature selection, using a chi-squared test, identifies key predictors. Feature scaling ensures consistent model training. The analysis employs various classification models, such as Logistic Regression, SVM, KNN, Naive Bayes, Decision Tree, Random Forest, and XG Boost. Model evaluation involves confusion matrices and accuracy scores, with visualizations for interpretation. This work contributes to advancing cervical cancer risk prediction methodologies.

4 Results

To employ machine learning algorithms for prediction of cervical cancer, we conducted an evaluation of accuracy and generated a confusion matrix as shown in figure 3 for each algorithm. Confusion matrix shows the actual details of how many how many instances were predicted correctly. We also calculated Accuracy, Precision, Recall and F-Measure from the values generated in confusion matrices.

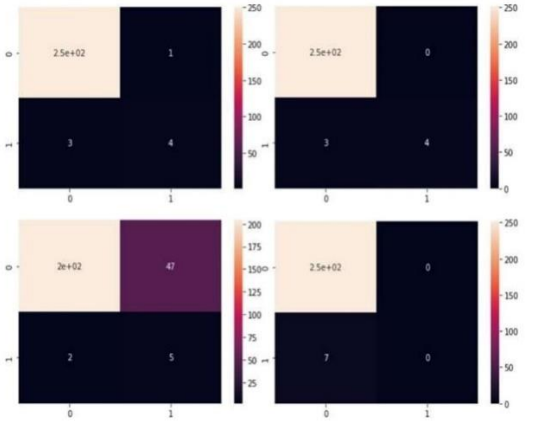

Fig. 3. Confusion Matrices of SVM, KNN, Naïve Bayes, Random Forest respectively

Following the assessment of the machine learning models' predictive accuracy for cervical cancer, we took the analysis a step further by visualizing the results on a table 1. This table 1 facilitated a straightforward comparison of various models' performance, aiding us in pinpointing the most accurate one.

Table 1. Comparison of Models

Sl. No	Name of the model	Accuracy%
1	Logistic Regression	98.44
2	Support Vector Machine	98.44
3	K-NN	98.83
4	Naïve Bayes	81
5	Decision Tree	99.22
6	Random Forest	97.28
7	XG Boost	99.22

By plotting the accuracy values of each model on the y-axis against the corresponding model names on the x-axis, we created a clear visual representation of their predictive abilities. The graph enabled us to observe the relative performance of the models at a glance, making it easier to identify the top-performing algorithms. The visual comparison given in figure 3 enabled us to gain insights into the machine learning models' effectiveness in predicting cervical cancer, primarily by evaluating their accuracy scores. We could determine which algorithms yielded the highest accuracy and make informed decisions about selecting the most suitable model for future analysis.

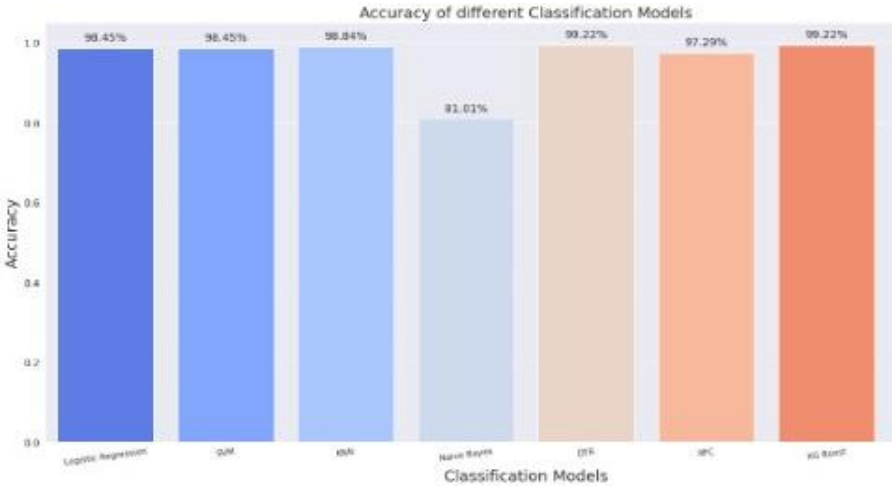

Fig. 4. Accuracy of different classification models

The paper presents and discusses the outcomes obtained by applying various machine learning methods to the "kag_risk_factors_cervical_cancer" dataset for cervical cancer classification. Performance metrics, encompassing accuracy, precision, recall, and the F1 score, are utilized to evaluate the model's effectiveness. The findings indicate that the machine learning model developed in this study effectively distinguishes cervical cancer cells. It achieves a high level of accuracy and maintains a satisfactory balance between precision and recall. This highlights the model's capability to capture

underlying cancer patterns through the selection of features and prioritization techniques. Moreover, these models showcase their efficacy and efficiency by identifying cases that were not encountered during the experiment. The favorable classification outcomes provide substantial evidence of machine learning techniques' potential to aid in cervical cancer diagnosis and risk assessment. Nevertheless, it is important to emphasize that further validation and testing on diverse and larger datasets are necessary to establish the model's reliability and generalizability. In conclusion, the results underscore the potential of machine learning to contribute to early detection and enhance clinical decision-making in cancer management.

5 Conclusion

In conclusion, the purpose of this paper is to harness data to classify cervical cancer using a range of machine learning approaches. The findings highlight the potential of machine learning methods in detecting and categorizing cervical cancer. Rigorous preprocessing, encompassing the handling of missing values, normalization, and feature engineering, readies the data for analysis. We construct a classification model using machine learning algorithms and assess their performance using appropriate metrics. The paper's findings shed light on critical factors related to cervical cancer, such as HPV infection, the count of sexual encounters and the age at initial sexual activity. Notably, our study reveals that the XG Boost model outperforms other machine learning algorithms in terms of accuracy. However, it is crucial to acknowledge the study's limitations. Our results are rooted in specific data and should undergo further validation with more extensive and diverse datasets to ensure the model's robustness and generalizability. Additionally, collaboration with medical professionals and specialists is imperative to correctly interpret the results and validate the model's outputs. In essence, this paper contributes to the growing body of research utilizing machine learning for cervical cancer classification. These results underscore the capacity of machine learning methods to improve cervical cancer diagnosis, assess risks, and inform decision-making. This paper lays the foundation for forthcoming research that can build upon these results, ultimately resulting in the creation of more effective and dependable models for cancer prediction and management.

References

1. Bosch FX, de Sanjose S.: The epidemiology of human papillomavirus infection and cervical cancer. Dis Markers. 23(4), 213-227(2007).
2. Saslow D, Solomon D, Lawson HW, et al.: American Cancer Society,American Society for Colposcopy and Cervical Pathology, and AmericanSociety for Clinical Pathology Screening Guidelines for the Prevention andEarly Detection of Cervical Cancer. Am J Clinical Pathology. 137(4), 516-542(2012).
3. Aryn M, Ronco G,Anttila A,et al.: Evidence regarding human papillomavirus testing in secondary prevention of cervical cancer. Vaccine. 30(Suppl5),F88-F99,(2012).

4. Arbyn M, Smith SB, Temin S, et al.: Detecting cervical precancer and reaching underscreened women by using HPV testing on self samples: updated meta-analyses. BMJ,363, k4823,(2018).
5. Schiffman M, Wentzensen N, Wacholder S, et al.: Human papillomavirus testing in the prevention of cervical cancer. J Natl Cancer Inst. 103(5),368-383, (2011).
6. Koh WJ, Abu-Rustum NR, Bean S, et al.: Cervical cancer, version 3.2020, NCCN Clinical Practice Guidelines in Oncology.J Natl Compr Cancer Network.18(6),660-682,(2020).
7. Chen W, Zheng R, Baade PD, et al.: Cancer statistics in China. CA Cancer Journal Clin. 66(2),115-132,(2016).
8. Franco EL, Villa LL, Sobrinho JP, et al.: Epidemiology of acquisition and clearance of cervical human papillomavirus infection in women from a high-risk area for cervical cancer. Journal Infect Dis. 180(5), 1415-1423,(1999).
9. Joura E A,Giuliano A R, Iversen O E,et al.:A9-valent HPV vaccine against infection and intraepithelial neoplasia in women.N Engl J Med. 372(8),711-723,2015.
10. Goldie SJ, Gaffikin L, Goldhaber-Fiebert JD, et al.: Cost- effectiveness ofcervical-cancer screening in five developing countries. N Engl J Med.353(20),2158-2168,(2005).
11. GauravKumawat,SantoshKumarVishwakarma,PrasunChakrabarti,Pankaj Chittora, Tulika Chakrabarti, and Jerry Chun-Wei Lin: Prognosisof Cervical Cancer Disease by Applying Machine Learning Techniques,Journal of Circuits, Systems and Computer,32(1), 2350019 (2023)
12. Kurman, S., Kisan, S.: An in-depth and contrasting survey of meta-heuristicapproaches with classical feature selection techniques specific to cervical cancer. Knowl Inf System, 65,1881–1934(2023).
13. Mamta Arora,Sanjeev Dhawan, Kulvinder Singh.: Cervical Cancer Diagnosis and Prediction: An Application of Machine Learning Techniques. Computational Intelligence in Analytics and InformationSystems.1-14, (2023).
14. Jiayi Lu, Enmin Song, Ahmed Ghoneim, Mubarak Alrashoud.: Machine learning for assisting cervical cancer diagnosis: An ensemble approach, Future Generation Computer Systems,106, 199-205,(2020).
15. N. Lavanya Devi & P. Thirumurugan.: Cervical CancerClassification from Pap Smear Images Using Modified Fuzzy C Means,PCA, and KNN, IETE Journal of Research, 68(3), 1591-1598,(2022).