Performance Evaluation of Supervised Learning Classifiers in Cervical Cancer Risk Stratification

[1]S. Deepa, [2].Dr. V. Priya
[1]*Department of Computer Science, Vellalar College For Women, Erode, Tamil Nadu, India*
[1]deepasamiappan@gmail.com, [2]priya@vcw.ac.in

ABSTRACT: Cervical cancer is a most important global health issue that affects a large number of women each year. It is one of the main causes of death in women around the world. Detecting cervical cancer earlier and accurately predicting the risk can greatly increase the chances of successful treatment and help save lives. This study aims to create a classification model using data mining techniques to identify people who are at high risk of developing cervical cancer. This enables timely medical action and treatment. The study uses data mining algorithms to determine which patients are at high risk of cervical cancer and whether they need a biopsy. After applying data preprocessing methods, several classification algorithms such as Naïve Bayes (NB), Instance-Based learner with parameter k (IBK), AdaBoost1, Random Forest (RF) and J48 were trained and tested. The performance of these models was evaluated using various classification measures like precision, recall and F-Measure. Among all the models, Random Forest showed the highest accuracy of 98.75% and it outperforms other algorithms.

Keywords: Cervical cancer, Naïve Bayes, IBK, Random Forest, J48, AdaBoost

1. INTRODUCTION:

Cervical cancer is commonly occurring cancer in women in recent days. It is causing approximately 280000 deaths every year and nearly 1 million women are suffering and living with cervical cancer. Nearly 85% of women suffering from cervical cancer is from under developing countries. Here are some countries which have the highest rate of cervical cancer such as Swaziland, Malawi, Zambia and Zimbabwe etc [1]. It is proved that if the cervical cancer is predicted in the early stage, it can be treated and cured. Cervical Cancer intrudes the deep tissues of the cervix and it has the ability to spread to other parts of the body such as the lungs, liver and vagina. This disease often progresses silently without any distinct symptoms at the early stage [2]. Since it does not show any early indicator, it can be diagnosed at the later stage which leads to increase in the treatment complexity. The study highlights the potential of data mining algorithms in identifying risks involved in cervical cancer detection. It suggests that advanced data mining techniques can be valuable tools in healthcare research and clinical applications. These methods simplify the identification of complex data patterns and it enables the extraction of hidden information and outcome prediction and thereby maximizing the value of the data. Various risk factors, patient data and Human Papilloma Virus (HPV) infection status are analyzed and these algorithms can provide risk assessments that help healthcare providers identify individuals at a higher risk of developing cervical cancer. Moreover, data mining techniques helps in increasing the accuracy of HPV risk assessments by considering multiple parameters and historical data [3].

This paper will provide an insight on the best classifier among NB, IBK, AdaBoost1, RF, J48 for classifying risk factor for cervical cancer. Section2 provides a literature review of works related to the predicting risks of cervical cancer using data mining techniques. Section 3 presents the classification methodology used to perform the data mining task along with the dataset. Section 4 provides the evaluation metrics and finally section 5 provides the conclusion.

2. LITERATURE REVIEW:

Cervical cancer remains a major public health issue and recognizing behavioral risk factors at an early stage can play a key role in lowering the number of cases and deaths from the disease. This section focuses on identifying cervical cancer through various risk factors and provides a general summary and in-depth review of the current research in this area.

With advancements in machine learning and web development, there is a growing interest in utilizing these technologies to improve cervical cancer prediction and visualization. Caruana and Niculescu conducted several studies that have shown the importance of ML algorithms in accurately predicting cervical cancer outcomes by leveraging various features such as demographic information, medical history and biomarkers to

train predictive models [4]. Various ML techniques, model evaluation and web development were employed to create a predictive model for cervical cancer. By applying advanced algorithms and providing an accessible platform, their research contributed to the early detection and prevention efforts which finally enhances awareness on cervical cancer and it assists in healthcare decision-making processes.

In2022, Naif and Abdulwahab proposed a machine learning-based model where the training data is initially provided to the system at the start of the model training process. Then machine learning algorithms are implemented on both the model input data and new input data and the scheme is made to properly train the architecture. Then predictions are made based on the newly accumulated data [5]. Various conventional machine learning principles and several traditional machine learning algorithms including Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) are used. For cervical cancer prediction, the highest classification score was achieved using the Random Forest (RF), Decision Tree (DT), Adaptive Boosting and Gradient Boosting algorithms. In contrast, the Support Vector Machine (SVM) achieved less accuracy than RF.

In 2025, Priyansh et al., discussed that cervical cancer accounts for 7.9% of cancers in women globally. It is a critical health challenge particularly due to its asymptomatic nature in early stages. They proposed a ML driven framework to predict cervical cancer risk using clinical data and address the class imbalance through SMOTE and optimizing model performance via GridSearchCV. By evaluating five classifiers—KNN, RF, LR, DT and Support Vector Classifier—this research identifies the DT model as the most robust, achieving superior precision, recall and accuracy metrics [7]. The data mining techniques are applied for enhancing early detection and personalized treatment strategies for cervical cancer.

In 2025, Sanjiv and Meenakshi sundaram suggested that cervical cancer remains a major public health problem and it needs accurate and interpretable risk prediction tools for diagnosis at an earlier stage. This study proposes a robust framework using an optimized Random Forest classifier for cervical cancer risk prediction, combined with Local Interpretable Model-agnostic Explanations (LIME). The RF model is optimized using key hyper parameters to enhance predictive accuracy. LIME provides local explanations, visualizing feature contributions for individual predictions, while global feature importance identifies the most influential risk factors [8]. The proposed framework achieves high accuracy and it shows a transparent decision-making and also enhances trust among medical professionals.

3. METHODOLOGY:

a. Dataset

The dataset is taken from UCI Machine Learning Repository[13]. The dataset focuses on risk factors associated with cervical cancer. The dataset contains a total of 858 patient records, encompassing 36 key features extracted from various demographic features, clinical data, lifestyle factors such as Age, Num_of_pregnancies, Smokes_years, IUD,STDs, STDs_HIV, STDs_Hepatitis_B, etc. Each entry represents an individual with specific data points. The features cover demographic information, habits and historic medical records. Among the total data, 800 records are taken for training and 58 records are taken for testing.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD
1										
2	18	4	15	1	0	0	0	0	0	0
3	15	1	14	1	0	0	0	0	0	0
4	34	1 ?		1	0	0	0	0	0	0
5	52	5	16	4	1	37	37	1	3	0
6	46	3	21	4	0	0	0	1	15	0
7	42	3	23	2	0	0	0	0	0	0
8	51	3	17	6	1	34	3.4	0	0	1
9	26	1	26	3	0	0	0	1	2	1
10	45	1	20	5	0	0	0	0	0	0
11	44	3	15 ?		1	1.266973	2.8	0	0 ?	
12	44	3	26	4	0	0	0	1	2	0
13	27	1	17	3	0	0	0	1	8	0
14	45	4	14	6	0	0	0	1	10	1

Fig.1 Dataset: risk_factors_cervical_cancer.csv

b.Algorithms

The previous work on literature review has showed several prominent algorithms for cervical cancer risk classification. This study intends to use five classification algorithms in data mining that are provide a comparative analysis using the provided dataset. The following classification algorithms are used for classification, **Naïve Bayes, IBK, AdaBoost1, Random Forest, J48.**

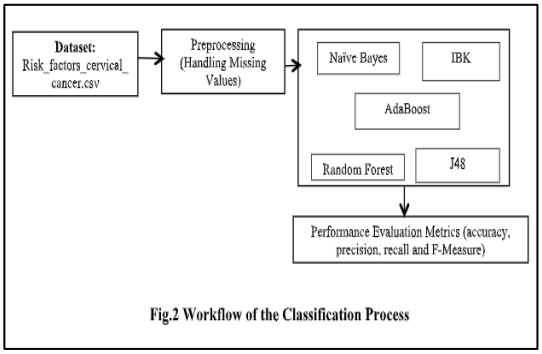

Fig.2 Workflow of the Classification Process

Naïve Bayes Algorithm (NB) :
NB is a supervised machine learning algorithm which is based on Bayes' theorem. It is referred to as "naïve" because it undertakes that all features in the dataset are independent of each other. The key idea behind this algorithm is that it calculates the probability that a data point belongs to a particular class based on prior knowledge (prior probability)and the likelihood of observing its features [14]. It is a graphical model based on Bayesian concept that has nodes corresponding to each of the columns or features. The Naïve Bayes classifier is a simple and powerful algorithm for solving classification problems. Though it works under the assumption of feature independence, it performs remarkably well for tasks like spam filtering, sentiment analysis and document classification. The algorithm is very effective with high dimensional data and works well even with small datasets.

IBK:
Instance Based K algorithm is able to standardize the range of features, incremental occurrence of processes (learned knowledge) and rules for inclusion of missing values. It uses distance metric that is similar to KNN but the search method applied is different from KNN [9]. The memory about training samples is effective in classification of new sample sand the algorithm uses distance metrics to predict new classes thereby reducing the training curve. It however maintains a large storage space for memorizing training cases that incurs high computational costs.

AdaBoost1:
AdaBoost is a powerful ensemble learning method that makes effective use of a small number of training examples by adjusting the importance of each example during the learning process. It is widely used for classification tasks. In the training process, the weight assigned to each sample is increased if the sample is classified incorrectly and decreased if it is classified correctly. This adjustment helps focus more on the examples that are difficult to classify. The algorithm continues to train new models based on these updated weights [10]. The goal is to improve the performance of subsequent models by minimizing their errors, ultimately leading to higher accuracy.

Random Forest Algorithm (RF):
RF is an ensemble supervised ML technique. Random Forest has tremendous potential of becoming a popular technique for future classifiers because its performance has been found to be comparable with ensemble techniques bagging and boosting. These ensemble methods aim at improving the classification accuracy by aggregating the predictions from multiple classifiers. RF algorithm uses two methods ie) Sub-sampling the examples/cases as in bagging and Sub-sampling the features known as feature selection [11]. These strategies are used in Random Forest to perform randomization and achieve diversity. There will be no pruning in the base decision trees which ensures diversity among them.

J48:
For classification and reduced error pruning this algorithm uses a greedy technique to induce DT. It is an ID3 extension. J48 calculates the result value of the new sample using predictive models and is based on various attribute values of available data [12]. The different properties are represented by the internal nodes of the decision tree, the branches between nodes indicate the possible values where attributes are included in the observed sample and the final value of the dependent variable is indicated in the terminal node.

4. PERFORMANCE ANALYSIS:

Dataset of 858 instances are taken for this comparative study and they are evaluated in terms of training and test dataset. The dataset is considered to have different details for each instance such as age, IUDs, STDs, other parameters that are needed to find out whether the patient has the risk of cervical cancer and they finally classify them as whether the patient needs to take biopsy or not. Various classification algorithms such as NB, IBK, AdaBoost, RF, J48 are taken and their performances are evaluated. The following table shows the total number of correctly classified and incorrectly classified instances among the training data and test data.

Table1: Classification results of training and test data

Algorithm	Total Number Of Instances		Number Of Correctly Classified Instances		Number Of Incorrectly Classified Instances	
	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data
Naive Bayes	800	58	720	52	80	6
IBK	800	58	796	56	4	2
AdaBoost	800	58	761	55	39	3
Random Forest	800	58	800	57	0	1
J48	800	58	781	55	19	3

From the above table, it is evident that among the five classification algorithms, RF performs the classification with more accuracy.

The following fig.3 and fig.4 shows the number of correctly classified and incorrectly classified instances in the training data and test data.

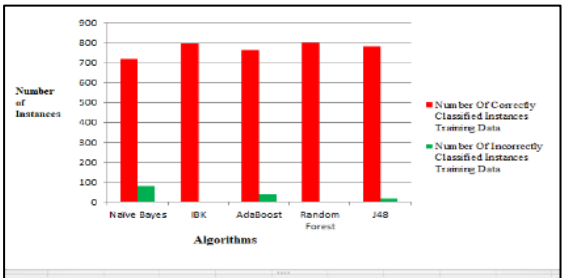

Fig.3. Classification of Training Data

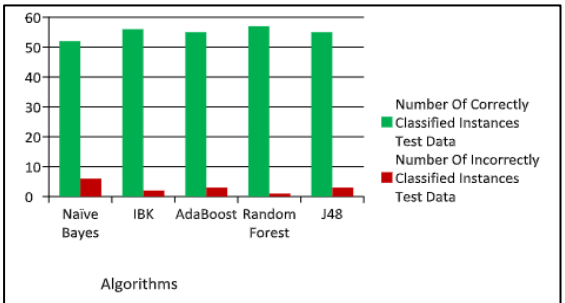

Fig.4. Classification of Test Data

Various evaluation metrics are used to assess the performance of the algorithms. The Kappa value indicates how effectively the classifier performed. The Mean Absolute Error (MAE) represents the average of the absolute differences between the predicted probabilities and the actual outcomes. The Root Mean Squared Error (RMSE) is similar to MAE, but it gives more weight to larger errors. The Relative Absolute Error (RAE) shows how much error the model produces compared to a simple baseline model, which always predicts the mean of the actual values. The Root Relative Squared Error (RRSE) is like RAE, but it is calculated using squared errors. Among the various evaluation metrics such as Kappa statistic, MAE, RMSE, RAE, and RRSE, the RF algorithm demonstrates the highest level of accuracy in classification.

Table 2. Evaluation Metrics of different parameters

Algorithm	Kappa Statistic		Mean Absolute Error		Root Mean Squared Error		Relative Absolute Error		Roor Relative Squared Error	
	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data
Naïve Bayes	0.4458	0.5246	0.1021	0.0948	0.2926	0.2886	84.843	75.9377	119.7757	113.8598
IBK	0.9565	0.6506	0.0062	0.0356	0.0706	0.1855	5.1595	28.5468	28.9113	73.1833
AdaBoost	0.5058	0.383	0.0531	0.0434	0.1719	0.165	44.157	34.7725	70.3556	65.0904
Random Forest	1	0.8482	0.024	0.044	0.0699	0.1479	19.9414	35.2308	28.6113	58.3645
J48	0.8028	0.383	0.0412	0.0634	0.1427	0.2244	34.2547	50.7632	58.3923	88.5367

The following table shows other performance evaluation factors such as precision, recall, F-measure, MCC, ROC Area and PRC Area. These values predict the quality and reliability of the classifier's predictions. Precision predicts how trustworthy the positive predictions are. Recall (Sensitivity) predicts how many actual positives the model can find. F-Measure predicts the balance between Precision and Recall. MCC value predicts the overall reliability of your classifier, considering all outcomes.

Table 3. Performance Evaluation Factors of different Algorithms

Algorithm	Precision		Recall		F-Measure		MCC		ROC Area		PRC Area	
	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data
Naïve Bayes	0.943	0.959	0.9	0.897	0.916	0.916	0.484	0.596	0.927	0.981	0.964	0.986
IBK	0.995	0.967	0.995	0.966	0.995	0.96	0.957	0.694	0.962	0.75	0.991	0.935
AdaBoost	0.945	0.951	0.951	0.948	0.946	0.933	0.521	0.487	0.979	1	0.977	1
Random Forest	1	0.983	1	0.983	1	0.982	1	0.858	1	0.995	1	0.996
J48	0.976	0.951	0.976	0.948	0.976	0.933	0.803	0.487	0.925	0.5	0.975	0.935

The following table shows the combined accuracy, precision, recall and F-Measure values of both training and test data. Among the five algorithms, RF shows the highest accuracy value. The precision, recall and F-Measure values are also better in RF than other four algorithms.

Table 4. Performance Evaluation Metrics

Algorithms	Accuracy (in %)	Precision (in %)	Recall (in %)	F-Measure (in %)
Naïve Bayes	89.98	0.944	0.9	0.916
IBK	99.3	0.993	0.993	0.994
AdaBoost	95.11	0.946	0.95	0.945
Random Forest	99.88	0.999	0.999	0.999
J48	97.44	0.998	0.974	0.987

The following figures shows the accuracy, precision, recall and F-measure values of different algorithms and here, RF provides better values for all the performance metrics.

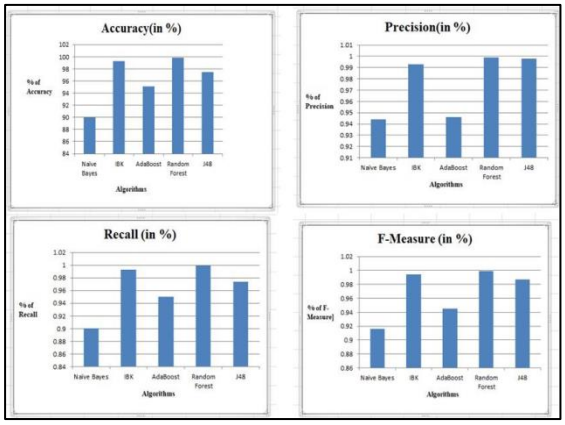

Fig.4.Classification of Test Data

5. CONCLUSION

Data Mining is a broad area that aims to solve many real world problems, especially in medical field. Classification could be a data processing technique supported ML which is employed to classify each item in a set of data into a group of predefined categories or teams. It encompasses a wide range of algorithms that solves variety of problems. These algorithms work by exploring the labeled training data to learn patterns and predict the category for new, unseen data points. In this work, five different classification algorithms have been used to predict whether a patient has a risk of cervical cancer and whether they are in need to take biopsy. Among the five different methods, RF proves to be more efficient in terms of classification.

References

1. M.S. Minu, "Cervical Cancer Prediction using Naïve Bayes Classification", International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

2. Jie Sue, et.al., "Cervical Cancer Prediction Using Machine Learning Models Based On Routine Blood Analysis", Scientific Reports volume15, Article number: 22655 (2025)

3. Ritu Chauhan, Anika Goel, Bhavya Alankar, "Predictive modeling and web-based tool for cervical cancer risk assessment: A comparative study of machine learning models", Volume 12, June 2024.

4. R. Caruana, A. Niculescu - Mizil, "An empirical comparison of supervised learning algorithms" Proceedings of the ACM International Conference Proceeding Series, 148 (2006), pp.161-168,10.1145/1143844.1143865

5. Naif Al Mudawi, Abdulwahab Alazeb, "A Model for Predicting Cervical Cancer Using Machine Learning Algorithms" , *Sensors*2022,*22*(11),4132;https://doi.org/10.3390/s22114132

6. J.J. Tanimu, M. Hamada, M. Hassan, H. Kakudi, J.O. Abiodun, "A machine learning method for classification of cervical cancer", Electronics 11 (3) (2022) 463

7. Priyansh Agarwal, Annanya Chaudhary, T.R.Sarvanan, " Cervical Cancer Risk Prediction Using Smote and Explainable AI", 2025 International Conference on Pervasive Computational Technologies (ICPCT),**DOI:** 10.1109/ICPCT64145.2025.10940701

8. Sajiv G; N Meenakshi sundaram, "Interpreting Cervical Cancer Risk Predictions Using Optimized Random Forest and Explainable AI Techniques", 2025 8thInternational Conference on Trends in Electronics and Informatics (ICOEI), **DOI:**10.1109/ICOEI65986.2025.11013521

9. Sivakami M, Dr. M. Thangaraj, P. Aruna Saraswathy, "A comparative analysis of machine learning techniques for automatic text classification", ISSN: 2454-132X, Volume 7, Issue 2 - V7I2-1523,2021.

10. Ender Sevinç, "An empowered AdaBoost algorithm implementation: A COVID-19 dataset study", Computers & Industrial Engineering165 (2022) 107912.

11. Vrushali Y Kulkarni, Dr Pradeep K Sinha, "Random Forest Classifiers :A Survey and Future Research Directions", International Journal of Advanced Computing, ISSN:2051-0845, Vol.36, Issue.1

12. Dr. Mutasim Al Sadig , Dr. Khalid Nazim, Abdul Sattar, "Developing a Prediction Model Using J48 Algorithm to Predict Symptoms of COVID-19 Causing Death", IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.8, August 2020

13. https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors.

14. Beslin Pajila, B. Gracelin Sheena, A. Gayathri. J. Aswin, "A Comprehensive Survey on Naïve Bayes Algorithm: Advantages, Limitations and Applications", Conference: 2023 4th International Conference on Smart Electronics and Communication (ICOSEC)