[1]Dr. Rashmi Ashtagi

Vaishali Rajput

Sonali Antad

Pratiksha Chopade

Atharva Chivate

Shreeshail Chitpur

Isha Dashetwar

Cervical Cancer Prediction Using Machine Learning

Abstract: - Cervical cancer constitutes a profound global health challenge, affecting a substantial number of women annually. If this cancer is not diagnosed and attended with hypervigilance, it can spread to other body parts, damage tissues, and typically deteriorate the immune system of the body. This ultimately becomes deadly and, in many cases, non-curable. This research paper looks into machine learning used to predict cervical cancer. The 5-year relative survival rate after the spread of the disease is almost 50%. Hence, detecting the tumor in advance can prevent its proliferation and consequently assist in the process of curing the disease.

The study emphasizes the global impact of cervical cancer, shedding light on its prevalence and its connection to the human papillomavirus (HPV). In a managerial but approachable manner, this research discusses the factors like weakened immune systems, smoking, and contraceptive use that contribute to the risk of cervical cancer. Four essential diagnostic tests—Hinselmann, Schiller, Citology, and Biopsy—are discussed as integral components of the predictive model. Diverse Machine Learning algorithms used show promise in enhancing accuracy for early cervical cancer prediction. With potential implications for public health, this paper concludes by providing insights into future research directions.

Keywords: Bagging Classifier, Cervical Cancer, Early Detection, HPV, Machine Learning, Predictive Modeling

I. INTRODUCTION

Cervical cancer is a significant global health issue that affects many women worldwide. This research paper looks into machine learning used to predict cervical cancer, offering a potential breakthrough in early detection. World Health Organization says that cervical cancer ranks 4th among cancers that affect women worldwide.

The rates of the female population being affected by cervical cancer have been increasing by 1.7% per year, primarily stemming from long-lasting infections with the Human Papillomavirus (HPV). While HPV is a major contributor, other factors can increase the chances of developing cervical cancer. Weakness in the defense system of the body, heavy smoking, prolonged use of birth control pills, numerous pregnancies, or exposure to certain medications during pregnancy, such as DES, can elevate the risk. The gravity of this disease is evident from the alarming numbers – approximately 570,000 new cases and 311,000 deaths are reported every year. Unfortunately, there is a lack of awareness regarding the statistics and associated risks of cervical cancer, making it imperative to address this information gap. This situation is more severe in developing and under-developed countries.

The tiny aperture of the uterus is called the cervix. It joins the birth canal to the upper portion of the uterus. The only gynecologic cancer that is consistently preventable with regular screening exams and necessary follow-up care is cervical cancer. In the US, 12,230 women are said to be diagnosed with Cervical Cancer annually; most of these cases affect women over 30. Cervical cancer is the third highest frequent disease in women globally. However, during the past 40 years, both occurrences and deaths have substantially decreased in the United States due to routine testing that detects precancerous cells early enough for treatment.

In the proposed research, machine learning is employed to make cervical cancer prediction more accurate and effective and facilitate the formation of preliminary predictions. The primary objective is to enhance preventive

[1]Department of Computer Engineering Vishwakarma Institute of Technology Pune, Maharashtra

rashmi.ashtagi@vit.edu

measures and enable timely interventions. Throughout this study, the aim is to fill the gaps in the understanding of cervical cancer, considering its implications for public health. This approach is conducted in a formal manner to ensure credibility and clarity in the findings. The paper seeks to investigate the causes of cervical cancer and its contributing factors comprehensively.

Simultaneously, the focus of the coursework extends to four diagnostic tests crucial to the predictive model: the Hinselmann Test, a colposcopic examination identifying abnormal blood vessels aiding early detection; the Schiller Test, using iodine to visualize abnormal cells indicating cervical cancer or precancerous lesions; the Citology Test or Pap smear, identifying cervix abnormalities through cell collection; and Biopsy Description, a definitive diagnostic procedure confirming cancer cell presence by examining a small cervix tissue sample. These tests collectively provide a comprehensive approach to screening and diagnosing cervical abnormalities. By incorporating these tests into our predictive model, we aim to improve the accuracy of early detection, allowing for timely medical interventions.

Predicting cervical cancer not only impacts individual health outcomes but also has broader implications for public health. The significance of predicting cervical cancer cannot be overstated. It is a complex health issue with multifaceted contributing factors, and machine learning offers a promising avenue to address this challenge. Through the research, the principal motivation is to contribute valuable insights that can inform healthcare practices, policy decisions, and public awareness initiatives, ultimately making strides toward a future with reduced cervical cancer incidence and improved women's health globally.

II. LITERATURE REVIEW

Cervical cancer is the predominant source of cancer-related deaths in women globally, which behaves epidemiologically like a low-infectious venereal illness. There are several risk factors for cervical cancer that are associated with HPV exposure. The implementation of screening also had an impact on the significant variations in incidence between nations. Liquid-based cytology (LBC), visual inspection with acetic acid and a typical Pap smear, and HPV testing are the main screening techniques used. However, the accuracy of the Pap smear and other techniques are not always accurate. Hence, there is a need to spread more awareness among women and society to undertake preventive measures. Cervical cancer can become the first cancer to be eliminated by humans [1].

HN Harsha Kumar et al. have proposed research and a case study. This case study aims to determine Indian women's awareness level about cervical cancer. Because cervical cancer is detected too late, a significant portion of cases have poor outcomes. 83 women from Mangalore City were selected at random and asked a variety of questions on cervical cancer. According to the statistics, just 7–8% of the general public is aware of cervical cancer and the process of getting screened for it. The majority of the women (85.5%) knew not much about the screening method, and 81.9% had limited awareness of this malignancy. As a result, the study concludes that there is an urgent need to inform doctors and women about cervical cancer, including its risks and treatments [2].

Riham Alsmariy et al. study utilized UCI's cervical cancer dataset, enhancing model performance with voting classification, SMOTE, PCA, and stratified 10-fold cross-validation, excelling in the Schiller test across multiple metrics. [3]. Matko Glučina's research revealed that using multilayer perceptron and K-nearest neighbors with oversampling methods like SMOTEEN and SMOTETOMEK significantly enhances early cervical cancer detection. Across all diagnostic methods, the strategy yielded over 0.95 in both mean AUC and MCC scores, indicating its effectiveness in early-stage cervical cancer diagnosis [4].

Sohely Jahan et al. focused on the early detection of cervical cancer. The study used many machine learning classification approaches based on risk markers to predict the probability of cervical cancer. The study compares the performance of several classification algorithms on the top five characteristics. The models were assessed for accuracy, precision, and recall, and the Multi-Layer Perceptron model consistently beat other techniques. In addition, the research addresses several dataset-splitting ratios, as opposed to prior publications that concentrated just on one [5]. Michał Kruczkowski et. al. used photonic methods to implement cervical cancer prediction using machine learning. This research study combined an optoelectronic sensor with various machine-learning algorithms to facilitate the early detection of cervical cancer. This approach involved four different algorithms: Random Forest, XGBoost, Naïve Bayes, and Convolutional Neural Networks (CNN), to discern between healthy and diseased tissue. The findings indicated high levels of accuracy, precision, recall, and F1-score across training datasets. Notably, the Naive Bayes classifier outperformed the others during the validation test, demonstrating an accuracy of 92%, a

precision of 93%, a recall rate of 93%, and an F1-score of 92%. Conversely, CNNs were less effective, potentially due to overfitting issues. In terms of both training and prediction speed, the Naive Bayes method proved to be the most efficient and accurate [6].

Breast cancer is one of the deadliest cancers and is a leading cause of female mortality. Hence, early detection is crucial for successful outcomes. The research proposes a deep learning model that integrates medial-lateral and craniocaudal perspectives and is effective for detecting breast cancer from computerized mammograms. For feature selection, the model relies on three distinct modules, and six unique categorization models are applied for diagnosis, achieving high efficiency and accuracy [7]. U N Wisesty et al. researched the gene mutation detection for breast cancer disease. The paper focuses on the early detection of breast cancer through the analysis of DNA abnormalities in blood cell samples, which is non-invasive and cost-effective. The paper also explores bioinformatics techniques based on DNA sequence data, which can be applied for breast cancer detection, including mapping the data, extracting features and prediction/classification using methods like Voss mapping, statistical feature representation, wavelet analysis, and regression approaches [8]. The article by X. Zhou et al. offers a thorough analysis of Artificial Neural Network (ANN)-based Breast Histopathology Image Analysis (BHIA) methods. It divides BHIA systems into deep and traditional neural networks and looks at pre-existing models to find relevant methods. The study talks about different ANN-based BHIA models. A CNN model based on DenseNet obtains 95.4% accuracy for classification tasks, whereas a model based on ResNet achieves a 98.77% correct classification rate [9].

Peng Xue et al. have provided a meta-analysis and a structured review that includes 20 studies to evaluate deep learning algorithms' diagnostic efficacy for early detection of cervical and breast cancer. The DL algorithms used have a pooled sensitivity of 88%, a specificity of 84%, and an AUC of 0.92, indicating acceptable diagnostic performance. The authors have suggested the need for evidence-based, standardized guidelines to raise the standard of DL research and prevent bias and overestimated performance of DL algorithms [10].

Jaswinder Singh et al. have developed a prototype for the prediction of cervical cancer. The research study involves the use of sensors applied to the fingertips, feet, and the area below the abdomen to the thighs to retrieve data on blood pressure, blood sugar level, and heart rate to scan for cervical cancer. The collected data is then validated against pre-defined contextual information. The Wi-Fi relay module ESP8266 is deployed for wireless communication and sending sensor data to the server's repository. The proposed prediction model for cervical cancer stages achieved accurate stage prediction in terms of false-positive rate, f-measure, and precision using machine learning classifiers. The results of the prediction model can be made available to physicians through REST channels, and authorized users can receive alerts in case of emergencies [11]. Naif Al Mudawi et al. have proposed research. The objective of the proposed study focuses on developing machine learning models for the prognosis of cervical cancer. There are four stages to it: dataset exploration, pre-processing of the data, choosing prediction models, and pseudo-coding. Making predictions and analyzing numerous machine learning models, such as KNN, Decision Tree, SVM, Random Forest, and many more, are part of the Predictive Model Selection section. However, the accuracy found is more than 99%, and in some instances, even 100%, indicating that the models are overfitting. In addition, a survey is carried out for the purpose of bringing people's attention to the human papillomavirus (HPV). According to this survey, much more work has to be done to raise public awareness of these kinds of illnesses [12]. However, the overfitting of the models can deteriorate the system.

Tanimu et al. conducted a survey and tested females for certain tests. The main intent of the study by Tanimu et. al. is to deploy machine learning frameworks for the prognosis of cervical cancer. There are four stages to it: dataset exploration, pre-processing of the data, choosing prediction models, and pseudo-coding. Making predictions and analysing various machine learning frameworks, such as KNN, Decision Tree, Random Forest, SVM and many more, are part of the Predictive Model Selection section. However, the accuracy found is more than 99%, and in some instances, even 100%, indicating that the models are overfitting. In addition, a survey is carried out to bring people's attention to the human papillomavirus (HPV). According to this survey, much more work has to be done to raise public awareness of these kinds of illnesses [13].

The number of females afflicted with cervical cancer is rising daily. There is a strong desire to stop this from happening using the tools available today. There is a tonne of untapped potential in the operation of spectroscopy and artificial intelligence for the identification of cervical cancer. Numerous spectroscopic methods, including mass spectrometry and fluorescence, have been used in cervical cytology. To enhance the cancer prediction system, several supervised and unsupervised learning algorithms can be introduced. Reduction of dimensionality approaches

such as PCA, data mining, and statistics techniques like HCA, SVM, RF, and numerous other company analyses may be included in this. Tremendously more research on neural networks can be done to improve these systems [14].

Milad Rahimi et al. conducted a descriptive investigation to integrate machine learning for predicting survival outcomes in cervical cancer patients. Their systematic analysis, based on 13 articles, revealed a diverse range of models employed, including Random Forest, SVM, logistic regression, ensemble, hybrid learning, and deep learning. Notably, the study identified varying sample sizes and validated models across the articles, with influential variables affecting survival prediction. Despite the evident advantages, challenges such as interpretability and imbalanced datasets persisted, emphasizing the necessity for further research to standardize ML algorithms for survival prediction in cervical cancer [15]. P. Rawt et al. focused on leveraging machine learning algorithms, which include bagging, logistic regression, XG Boost, and Random Forest, to detect cervical cancer in the early stage. Addressing difficulties such as imbalanced datasets, the study proposed a combined approach to enhance accuracy, leading to improved predictive accuracy and precision compared to individual algorithms. The research underscored the importance of machine learning in advancing of detection methods for Cervical Cancer, thereby contributing to enhanced treatment outcomes. Additionally, it emphasized the need for continued research in this domain [16]. P. Chakrabarti et al. examine the impact of gender expectations on women's cancer screenings; this study utilises machine learning, focusing on breast and cervical cancers. It explores prevalent women's health issues and emphasizes the role of deep learning, employing Convolutional Neural Networks with transfer learning (VGG16), to enhance early detection accuracy. The objective is to reduce mortality rates through advanced and timely predictions, shedding light on the crucial intersection of gender expectations and medical decision-making in women's cancer diagnostics.[17]

A. Arora et al. explored the pervasive occurrence of cervical cancer in the female population, emphasizing the importance of early identification through modalities such as pap-smear tests and colposcopy. Employing contour models that are active with Gaussian fitting energy for image segmentation, the study achieved a commendable 92% concordance with manually annotated images provided by expert cytologists. Particularly noteworthy was the utilization of polynomial SVMs, resulting in an impressive 95% accuracy in cell classification. This highlighted the potential of the proposed approach for precise and efficient cervical cancer diagnosis [18].

III. METHODOLOGY

The research methodology involves analyzing a dataset that includes results from these tests. A thorough overview is provided on the implementation process, normalization techniques, and application of seven machine learning algorithms, including K-Nearest Neighbor and Support Vector Machine, among many others.

1.1 About the Dataset

The dataset used in this research work is a Kaggle and includes 36 variables that indicate cervical cancer risk. These characteristics are collected from medical tests used to identify clinical findings associated with cervical cancer. The collection contains a wide range of information, including smoking habits, sexual behavior, and medical examination findings. To solve the issue of missing values in this dataset, numerous strategies, such as using mean and median values, have been used. Key characteristics in this dataset that play an important role in predicting cervical cancer include the use of contraceptive pills, alcohol intake, the number of sexual partners, and other physiological parameters. The mean values of some of the important parameters are as follows:

Feature	Mean Value
Age	26.82051282
Number of sexual partners	2.527644231
First sexual intercourse	16.99529965
Number of pregnancies	2.275561097
Smokes (years)	1.219721413
Hormonal Contraceptives (years)	2.256419201
IUD (years)	0.5148043185
STDs (number)	0.176626826

Table 1. Mean values of certain important parameters.

IUD stands for intrauterine device as a contraceptive. Some research suggests that the IUD may have a protective effect against cervical cancer. The immune response triggered by the IUD may help prevent cervical cancer. The above values suggest that the younger population is majorly infected by this tumor, which when not diagnosed timely can be deadly. The graph of age distribution clearly indicates that the women of age group 16 - 26 years are hit hard by this cancer, and the inducement is because of the usage of contraceptives, having multiple intercourse partners, and smoking. This ultimately causes the spread of Sexually Transmitted Diseases (STDs).

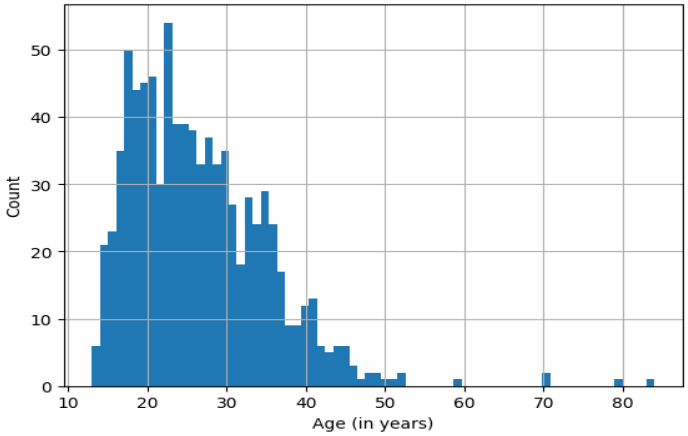

Fig. 1. Age distribution of affected population

This dataset has been further explored, pre-processed and later different machine learning algorithms have been deployed onto it.

1.2 Algorithms Used

KNN

Known for its simplicity and effectiveness, K-Nearest Neighbours (KNN) is a neighbourhood classifier that is frequently used in pattern recognition and machine learning. By allocating the label based on the majority of examples among its k-nearest neighbours in the training dataset, it classifies unlabeled test cases. The choice of distance metric significantly influences its performance. Although KNN is efficient, because it relies on each example in the training set, it has memory needs and time complexity issues. KNN is used in classification contexts where each category's attributes are pre-studied, the closest k class is identified using distance metrics, and test data is assigned to the k-nearest neighbour group that has the greatest number of members in a particular class. To determine the ideal k number, experiments are frequently carried out, and computations using the Euclidean distance are frequently made.

SVM

Designed for two-group classification tasks, the Support Vector Machine (SVM) is a supervised learning method. The process creates a linear decision surface with unique properties that guarantee a high degree of generalization ability by mapping input vectors to a high-dimensional feature space. SVM is a non-probabilistic binary linear classifier that, given a training set, builds a model to categorize testing set elements. It efficiently handles both linear and non-linear classifications using the kernel trick, with options like cubic and Gaussian SVMs achieved through different kernel functions.

Decision Tree

A decision tree serves as a classifier by dividing the input space into smaller segments and assigning labels to these segments according to different output categories. While traditional decision trees only partition along coordinate axes, they can grow to recognize subtle patterns. However, overgrown trees may lead to overfitting. Using decision rules and analytical features, this unsupervised learning technique builds prediction models for regression and classification tasks.

Random Forest

Random Forest (RF) is an ensemble learning method that utilizes decision trees for both regression and classification tasks. It employs a bootstrap specimen size for training each tree and selects optimal separation factors from a randomly chosen subset of all elements. For regression, the algorithm uses variance decrease, while for classification, it employs the Gini coefficient. RF combines the predictions through a majority of votes or an average. Known for their accuracy, robustness, and ease of use, Random Forest models are widely used in machine learning.

ADA Boost

ADA Boost starts with all training observations having equal weights and then utilizes a sequence of weak models. Notably, it increases the weight of misclassified observations, focusing on correcting errors. By integrating the effects of decision boundaries obtained from multiple iterations and combining results from several weak models, ADA Boost enhances the accuracy of misclassified findings and overall iteration outcomes. This iterative approach allows ADA Boost to effectively improve the model's performance by learning from its mistakes and refining decision boundaries.

Voting Classifier

A voting classifier is a proposed machine learning process that applies a variety of models to an ensemble and produces an output or class depending on the outcome of their combined efforts, that is, the output that has the highest probability predicted for the selected class. Instead of modelling different algorithms and finding their accuracies, the voting classifier creates one model which is trained by different models and predicts results based on the total predominance of the result of each model. The voting classifier can be further divided into two types of voting: these include a. Hard Voting and b. Soft Voting.

Bagging Classifier

Bagging, short for Bootstrap Aggregating, is a method where multiple subsets of the original dataset are used to build separate models. These models are then combined to form a final prediction that is more robust and less prone to overfitting. It is a technique that harnesses the collective wisdom of multiple models to increase accuracy and stability. In simple terms, bagging creates an ensemble of models, each trained on a different subset of the data, and then aggregates their predictions for enhanced accuracy and robustness.

The above-mentioned machine learning algorithms have been implemented, and seven different machine learning models have been developed, trained and tested for unknown values from the dataset. Several observations have been scrutinized from the models and the accuracies their accuracies.

1.3 Implementation

The main motive for the implementation of the system is to obtain more accurate and precise results for cervical cancer prediction. The dataset included the findings from the Hinselmann, Schiller, Citology, and Biopsy tests for cervical cancer. These four tests yield results that are used to build and optimize a number of machine-learning models. Prior to deployment, an analysis revealed that harmonical contraceptives were the primary cause of the tumor. It was discovered that women who regularly use harmonical contraception are more likely to get cervical cancer. Consequently, the harmonical contraceptives column was given precedence during the feature extraction process. Furthermore, the study revealed that women who smoke more frequently, consume more cigarettes annually, or have multiple intercourse partners may also have a more noticeable tumor. Therefore, these three factors were more emphasized over the others, that is, the harmonical contraceptives, number of intercourse partners, and the smokes (year). These three factors suggest that the female population needs to be educated about the usage of contraceptives and maintaining proper social relations with the surrounding.

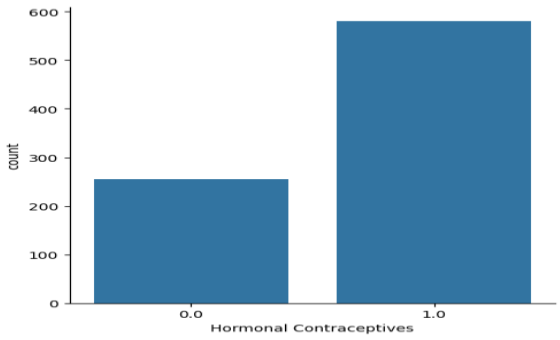

Figure. 2. Consumption of harmonical contraceptives versus count of women affected by cervical cancer.

Fig. 2. advocates that women who use contraceptives are more prone to be seized by this disease. The highest positivity rate of cervical cancer is found to be in the population with the highest usage of harmonical contraceptives. Fig. 3. displays the rate of positivity for four different cervical cancer tests against the usage of harmonical contraceptives. These figures clearly suggest that the improper and excessive usage of contraceptives is a major concern for the spread of this cancer.

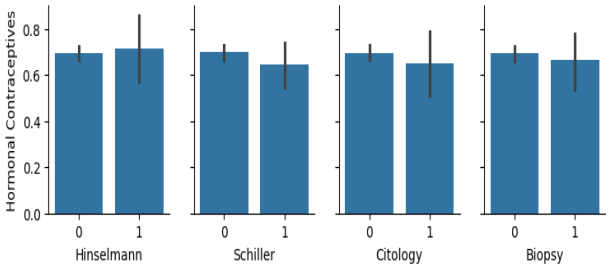

Figure 3. Graphs depicting the results of the four tests of the women with the consumption of harmonical contraceptives

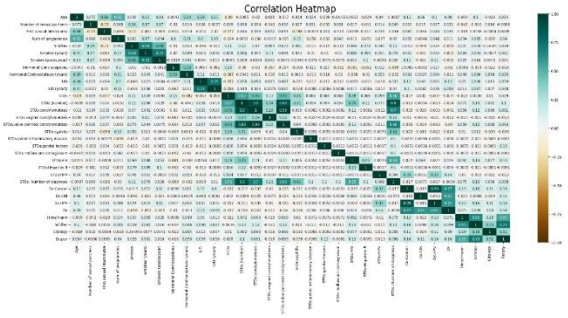

Figure. 4. Correlation HeatMap of the dataset.

After the determination of the important key features, the entire dataset was normalized. Normalization is basically a scaling technique in machine learning that is imposed during the pre-processing of the data, such that the complete dataset can be upscaled or downscaled to a common metric. The value of normalization can be mathematically calculated as:

$$Xn = \frac{(X - Xmin)}{(Xmax - Xmin)}$$

where,

X_n = ‘value of normalization’

$X_{maximum}$ = ‘maximum value of a feature’

$X_{minimum}$ = ‘minimum value of a feature’

Additionally, the null values were checked, and outliers existing in the dataset were also withdrawn. The dataset was completely pre-processed before utilization. The dataset was split into a 75-25% ratio, which implies 75% of the random dataset will be used for training purposes and the other 25% of the dataset will be used for testing the models and making predictions.

Altogether, seven machine learning algorithms were implemented and seven different models were developed, which include K-Nearest Neighbors, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, Ada Boost, Voting Classifier, and Bagging Classifier. Initially, the K-Nearest Neighbors algorithm was implemented. The parameter for the KNeighborsClassifiers was set to 5, indicating that the model will consider the label of the five nearest neighbors while making predictions.

In SVM, a polynomial kernel was imposed. A polynomial kernel is usually used to capture nonlinear relationships in the data. The regularization parameter, with a value of 1, controls the trade-off between having a smooth decision boundary and correctly classifying the training points.

In Decision Tree Classification, the class weight is set to none, indicating that all the classes have equal weights and no there is no imbalance in the dataset. The criterion parameter defines the function that is utilized in the decision tree to evaluate the split's quality. The term gini denotes the Gini impurity, which quantifies the likelihood of an element selected at random being mislabeled. The maximum depth of the tree is set to 2. Hence, an instance of the Decision Tree Classifier is set up with a maximum depth of 2, no class weights, and Gini impurity as the splitting criterion. A Random Forest is an ensemble learning technique that generates the class that is the mode of the classes (classification) by building an extensive amount of decision trees during training. Here, in this case, the value of n_estimators is declared to 100; therefore, 100 decision trees will be produced, and based on the mode value of the classes, the output will be generated.

Ada Boost builds a strong classifier by iteratively assigning greater weight to instances that are incorrectly categorized using the predictions of weak learners—in this instance, decision trees. Here, the maximum depth value is assigned to 500. The number of weak learners that will form the final boosted model is kept at 50. Hence, it is very much expected that this model will produce more accurate results.

A voting classifier is a technique for ensemble learning that aggregates the predictions from multiple independent classifiers. SVM, Decision Tree, Random Forest, and AdaBoost are the four pre-instantiated classifiers. All these classifiers are combined. A majority vote between the individual classifiers for classification tasks will determine the ensemble's final estimate. Comparing this method to individual classifiers, better generalization performance is frequently achieved.

Bootstrap Aggregating, also called Bagging Classifier, has been built and deployed on the dataset and is expected to provide the most accurate and authenticated decisions. As in bagging, unconnected sections from the training dataset are used for training several instances of the same base estimator, and the predictions of these instances are then combined. Here, an already instantiated instance of a decision tree classifier is used as a base estimator. The n estimator parameter is set to 10. Therefore, ten decision trees will be trained, and a random seed value will be chosen.

Seven machine learning models in all have been evaluated and trained, and their accuracy and precision values have been computed. The procedure of comparing the algorithms' performances under various conditions is made more facile by all of these models. In addition to facilitating early cervical cancer detection, they can be advantageous when formulating the most accurate predictions.

IV. RESULTS AND DISCUSSIONS

Early cervical cancer prediction was successfully implemented. Various Machine Learning models were deployed and tested on random unknown values. The parameters considered for evaluation are accuracy, precision, recall, and f1-score. These parameters are computed based on the predictions the model has made.

$$\text{Accuracy} = TP + TN/TP + TN + FP + FN$$

$$\text{Precision} = TP/TP + FP$$

$$\text{Recall} = TP/TP + FN$$

f1-score = precision * recall/precision + recall

TP = ‘true positive’

TN = ‘true negative’

FP = ‘false positive’

FN = ‘false negative’

Using the above parameters, all the models were evaluated. The outcomes of the different models are as follows:

Table 2. KNN Classification Model Report

	precision	recall	f1-score	support
0	0.95	0.98	0.97	193
1	0.64	0.44	0.52	16
accuracy			0.94	209
macro avg	0.8	0.71	0.74	209
weighted avg	0.93	0.94	0.93	209

Table 3. SVM Classification Model Report

	precision	recall	f1-score	support
0	0.96	0.97	0.97	193
1	0.6	0.56	0.58	16
accuracy			0.94	209
macro avg	0.78	0.77	0.77	209
weighted avg	0.94	0.94	0.94	209

Table 4. Decision Tree Classification Model Report

	precision	recall	f1-score	support
0	0.99	0.98	0.98	193
1	0.78	0.88	0.82	16
accuracy			0.97	209
macro avg	0.88	0.93	0.9	209
weighted avg	0.97	0.97	0.97	209

Table 5. Random Forest Classification Model Report

	precision	recall	f1-score	support
0	0.97	0.98	0.97	193
1	0.71	0.62	0.67	16
accuracy			0.95	209
macro avg	0.84	0.8	0.82	209
weighted avg	0.95	0.95	0.95	209

Table 6. AdaBoost Classification Model Report

	precision	recall	f1-score	support
0	0.96	0.98	0.97	193
1	0.73	0.5	0.59	16
accuracy			0.95	209
macro avg	0.84	0.74	0.78	209
weighted avg	0.94	0.95	0.94	209

Table 7. Voting Classification Model Report

	precision	recall	f1-score	support
0	0.96	0.98	0.97	193
1	0.69	0.56	0.62	16
accuracy			0.95	209
macro avg	0.83	0.77	0.8	209
weighted avg	0.94	0.95	0.94	209

Table 8. Bagging Classification Model Report

	precision	recall	f1-score	support
0	0.98	0.98	0.98	193
1	0.8	0.75	0.77	16
accuracy			0.97	209
macro avg	0.89	0.87	0.88	209
weighted avg	0.97	0.97	0.97	209

From Table 2, 3, 4, 5, 6, 7, and 8, it can be concluded that the decision tree provides quite satisfactory results with an accuracy of 97%. As the Bagging Classifier uses decision tree as its base estimator, the accuracy of Bagging Classification model is also 97%.

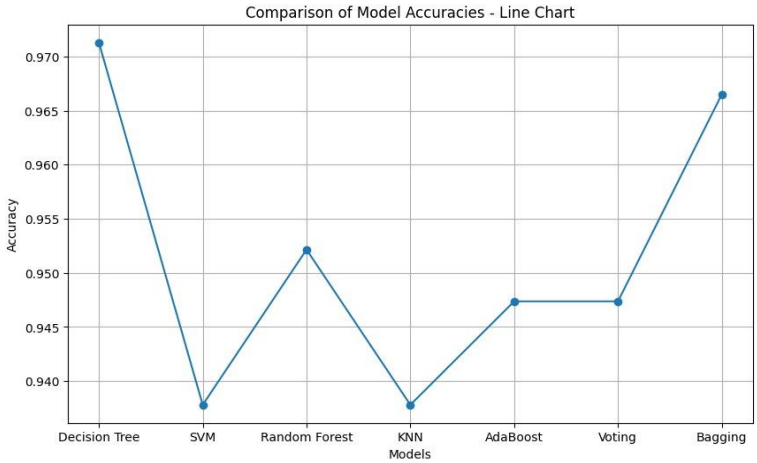

Figure 5. Comparison of seven machine learning models

The comparative analysis shows that the decision tree classifier has the highest accuracy. However, as the Bagging classification trains the decision tree classifier more rigorously, the bagging provides the most accurate outcomes

and also avoids any overfitting or underfitting. This comparative analysis indicates that the Bagging Algorithm can be more prominently used for the detection of cervical cancer at its embryonic stage.

V. CONCLUSION

Predicting cervical cancer holds immense importance in the realm of healthcare for women. Early detection significantly increases the chances of successful treatment, thereby reducing mortality rates. Machine learning plays a pivotal role in this scenario by analyzing vast datasets, identifying patterns, and providing predictive insights. The integration of technology in healthcare, particularly in the field of cervical cancer prediction, marks a crucial advancement that has the potential to revolutionize women's healthcare globally.

The proposed study suggests that the Bagging Classification Algorithm is the most suitable for early cervical cancer detection. This early detection can facilitate ushering up the treatment for this disease. Additionally, it must be considered that mere early prediction cannot help in reducing the number of infected women. Hence, the female population needs to be made acquainted with cervical cancer and its causes, repercussions, and treatment. There is a pressing urge to spread awareness about this cancer and suggest the precautions to be considered in everyday life to circumvent cervical cancer.

Through the identification of at-risk groups and the application of focused preventive strategies, there is the potential to alleviate the healthcare system's overall burden associated with cervical cancer. This approach aligns with global efforts to enhance preventive healthcare strategies and improve overall health outcomes.

VI. FUTURE SCOPE

The research opens up new possibilities for future exploration in the field of detecting cervical cancer. The study uses various algorithms of machine learning to predict cervical cancer. This domain can be further probed in various dimensions, which can include:

1. The research can further explore deep learning algorithms such as convolutional neural networks to analyze medical images and identify early signs, enabling the prediction of cervical cancer even earlier.

2. The potential of wearable devices and mobile health applications can be investigated to collect real-time data on lifestyle factors and symptoms, enabling personalized risk assessment and early intervention.

3. Large-scale population studies can be conducted to validate the predictive model across diverse demographics and geographical regions to ensure its effectiveness for different populations. Furthermore, collaborating with healthcare providers and policymakers to integrate the predictive model into routine screening programs enables proactive identification of high-risk individuals and targeted interventions.

4. Investigating the application of natural language processing (NLP) techniques in the analysis of electronic health records, with the aim of extracting pertinent information for the prediction of cervical cancer.

5. In addition, other factors, such as genetic data and biomarkers, can be incorporated to improve accuracy and early detection.

REFERENCES

[1] Zhang S, Xu H, Zhang L, Qiao Y. Cervical cancer: Epidemiology, risk factors and screening. Chin J Cancer Res. 2020 Dec 31;32(6):720-728. doi: 10.21147/j.issn.1000-9604.2020.06.05. PMID: 33446995; PMCID: PMC7797226.

[2] Harsha Kumar H, Tanya S. A Study on Knowledge and Screening for Cervical Cancer among Women in Mangalore City. Ann Med Health Sci Res. 2014 Sep;4(5):751-6. doi: 10.4103/2141-9248.141547. PMID: 25328788; PMCID: PMC4199169.

[3] Riham Alsmariy, Graham Healy and Hoda Abdelhafez, “Predicting Cervical Cancer using Machine Learning Methods” International Journal of Advanced Computer Science and Applications(IJACSA), 11(7), 2020.

[4] Glučina, Matko, Ariana Lorencin, Nikola Anđelić, and Ivan Lorencin. 2023. "Cervical Cancer Diagnostics Using Machine Learning Algorithms and Class Balancing Techniques" Applied Sciences 13, no. 2: 1061.

[5] Jahan, Sohely & Islam, Manowarul & Islam, Linta & Rashme, Tamanna & Prova, Ayesha & Paul, Bikash Kumar & Islam, M. & Mosharof, Mohammed. (2021). Automated invasive cervical cancer disease detection at early stage through suitable machine learning model. SN Applied Sciences. 3. 10.1007/s42452-021-04786-z.

[6] Kruczkowski, M., Drabik-Kruczkowska, A., Marciniak, A. et al. Predictions of cervical cancer identification by photonic method combined with machine learning. Sci Rep 12, 3762 (2022).

[7] Khalid, A.; Mehmood, A.; Alabrah, A.; Alkhamees, B.F.; Amin, F.; AlSalman, H.; Choi, G.S. Breast Cancer Detection and Prevention Using Machine Learning. Diagnostics 2023, 13, 3113.

[8] U N Wisesty et al 2020 IOP Conf. Ser.: Mater. Sci. Eng. 830 032051 DOI 10.1088/1757-899X/830/3/032051

[9] X. Zhou et al., "A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks," in IEEE Access, vol. 8, pp. 90931-90956, 2020, doi: 10.1109/ACCESS.2020.2993788.

[10] Xue, P., Wang, J., Qin, D. et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. npj Digit. Med. 5, 19 (2022).

[11] Singh, Jaswinder and Sandeep Sharma. “Prediction of Cervical Cancer Using Machine Learning Techniques.” (2019).

[12] Al Mudawi N, Alazeb A. A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. Sensors (Basel). 2022 May 29;22(11):4132. doi: 10.3390/s22114132. PMID: 35684753; PMCID: PMC9185380.

[13] Tanimu, Jesse Jeremiah, Mohamed Hamada, Mohammed Hassan, Habeebah Kakudi, and John Oladunjoye Abiodun. 2022. "A Machine Learning Method for Classification of Cervical Cancer" Electronics 11, no. 3: 463.

[14] Carlos A. Meza Ramirez, Michael Greenop, Yasser A. Almoshawah, Pierre L. Martin Hirsch and Ihtesham U. Rehman, Advancing cervical cancer diagnosis and screening with spectroscopy and machine learning, doi.: 10.1080/14737159.2023.2203816, 2023

[15] Rahimi, M., Akbari, A., Asadi, F. et al. Cervical cancer survival prediction by machine learning algorithms: a systematic review. BMC Cancer 23, 341 (2023).

[16] P. Rawat, M. Bajaj, S. Mehta, V. Sharma and S. Vats, "A Study on Cervical Cancer Prediction using Various Machine Learning Approaches," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 1101-1107, doi: 10.1109/ICIDCA56705.2023.10099493.

[17] G. S. P. Ghantasala, B. T. Hung and P. Chakrabarti, "An Approach For Cervical and Breast Cancer Classification Using Deep Learning: A Comprehensive Survey," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICCCI56745.2023.10128454.

[18] Arora, A. Tripathi and A. Bhan, "Classification of Cervical Cancer Detection using Machine Learning Algorithms," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 827-835, doi: 10.1109/ICICT50816.2021.9358570.