Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

A Machine Learning Pipeline for Cervical Cancer Prediction

Yubo Huang *

Henan University, Jinming College Jinming Avenue North Section, Longting District, Kaifeng City, Henan Province, China

* Corresponding Author Email: huangyubo@henu.edu.cn

Abstract. Cervical cancer ranks as the fourth leading cause of cancer-related deaths among women and often remains asymptomatic during its initial stages. Therefore, it's very important to predict it. The study's objective is to explore cervical cancer risk variables and develop an ensemble prediction model to forecast cervical cancer more precisely. The cervical cancer data were represented by 11 risk factors and a dependent variable consisted of Hinselmann, Schiller, Cytology, and Biopsy. SVM, decision tree, and random forest methods are compared for prediction, with models trained on 75% of the data and evaluated using confusion matrices. Precision and recall metrics assess model performance, with recall being vital in critical situations like identifying cervical cancer patients. The comparison was made among these three methods and compared the ranking result of risk factors with the ground truth, random forest method offers high recall (0.8821) and accuracy (0.8744). Recognizing risk factors can enhance cure rates and female patient survival. And the cervical cancer prediction system enhances medical care and reduces the cost. Data mining techniques in the medical sector facilitate diagnostic and prognostic apps for early treatment of life-threatening diseases.

Keywords: Cervical cancer; Machine learning; SVM; Decision tree; Random Forest.

1. Introduction

Cervical cancer originates in a woman's cervix, the gateway from the vagina to the uterus. In 2018, women who were diagnosed with cervical cancer are about 570000, and approximately 311000 women died from the disease [1]. Machine learning (ML) and deep learning are being used by more and more people and businesses to evaluate vast volumes of data and provide useful insights. In medical practice, it is more common to use ML-based techniques to predict the early stages of major illnesses including cancer, renal failure, and heart attacks [2]. In recent times, a considerable amount of research has been devoted to cervical cancer, employing advanced methodologies that enable early-stage prediction. The adoption of advanced machine learning techniques has played an important role in driving these predictive capabilities [3]. The integration of machine learning has not only improved research efficiency but has also enhanced the precision of patient data. However, specific regions face challenges such as limited awareness, restricted medical access, and financial constraints, which collectively contribute to the prevalence of cervical cancer among female populations [3]. Moreover, the incorporation of machine learning has not only bolstered predictive accuracy but has also unlocked innovative applications.

Earlier researchers have used a variety of classic ML-based strategies, such as k-nearest neighbors (KNN), K-means clustering, and random forest, to diagnose cervical cancer [4]. In Ashok et al. research, the SVM, RF, KNN, GLCM, and hierarchical cluster algorithms were discussed. In 2016, Vidya et al. made significant advancements in predicting typical cervix characteristics by harnessing the power of effective data mining tools to assess cervical cancer [5]. In 2019, the unequal distribution of data and risk variables for cervical cancer diagnosis was discussed by Geetha et al. [6]. Alyafeai et al. developed a fully integrated cervical image-based cervical cancer detection and screening pipeline in 2020. Currently, two deep neural network learning models are in development for the automated diagnosis and identification of cervical tumors [7].

In order to effectively reduce the risk of cervical cancer, a machine learning algorithm that can detect cancer as soon as possible is created. The dataset was used to predict whether a patient can have a cervical cancer or not. Different from previous studies, several models are used to predict cervical cancer. By comparing their performance, random forest was chosen to predict cervical cancer.

2. Methods

2.1. Data preprocessing

The dataset, collected from UCI, is related to Cervical Cancer and its prediction factors. The dataset contains numerous instances of missing values. The absence of these values can suggest various underlying distinctions. Instances featuring missing values can be incorporated, excluded, or replaced by either the mean value (for numerical attributes) or the most prevalent value (for categorical attributes). By implementing the method of eliminating rows with missing data, missing values can replace with median values. Our objective is to standardize the count of attributes while maintaining the count of input vectors in the dataset.

There are 35 variables in this dataset. 11 variables are continuous, such as age, number of sexual partners and number of pregnancies. 24 variables are binary, such as smokes, hormonal contraceptives and STDs.

Multicollinearity among independent variables leads to less reliable statistical judgements. And the standard errors of one or more variables in the model are inflated. Therefore, the variables with multicollinearity in the dataset should be removed. The findings of the cervical examinations are directly affected by representation of "Hinselmann," "Schiller," "Cytology," and "Biopsy". Then, a variable was created and defined as "Cervical Cancer" that is formed by the four factors. "Hinselmann," "Schiller," "Cytology," and "Biopsy" are all Binary Variables. The more of the values, the more risk of having cervical cancer, but a positive result does not always signify that the patient has the disease. For cervical cancer detection result, the ‘Cervical Cancer’ values greater than 1 are considered as people who have cervical cancer and the values not greater than 1 are considered as negative outcomes.

2.2. Feature selection

Feature selection is a critical step in the process of preparing data for analysis and modeling. Including too many features that are not truly informative can cause the model to fit noise in the training data. Feature selection helps to mitigate overfitting by focusing on the most pertinent features. Meanwhile, models with fewer features generally require less computational resources for both training and making predictions. This can lead to faster model development and deployment [8].

The Boruta method evaluates the significance of features by introducing randomness through the shuffling of predictor values. These shuffled values are then combined with the original predictors to create a merged dataset. A random forest model is constructed using this merged dataset. Subsequently, a comparison is made between the importance of the actual variables and that of the randomized variables. Variables with importance ratios higher than those of their randomized counterparts are selected. The Boruta method is often preferred over other feature selection packages in R due to its array of advantages and benefits. For this dataset, 11 variables are selected by using Boruta algorithm. With this exercise, the Boruta method begins by making a replica of the original dataset while removing the columns for the medical outcomes. Next, the Boruta method was used after setting the seed. Then it confirmed 10, eliminated 8, and 3 are put as tentative. Then, using internal function in the context of the Boruta algorithm serves the purpose of refining the status of "tentative" features. These "tentative" features are those that lack a clear importance decision in the initial Boruta analysis and require further processing to determine their final status.

2.3. Model Comparison

2.3.1. Support Vector Machine (SVM).

Support Vector Machine [9] is a supervised learning algorithm primarily designed for classification. While it can also manage regression, its strength primarily lies in classifying data. SVM's main goal is to determine the best-fitting hyperplane in a multidimensional space that distinctly separates data points into their respective classes. This hyperplane is optimized to ensure the widest possible margin between the nearest data points from different classes. The dimensionality of the hyperplane is influenced by the number of input features. As shown in figure 1, when they are two features, it's a line; for three, it's a 2D plane. Visualizing this becomes challenging when there are more than three features.

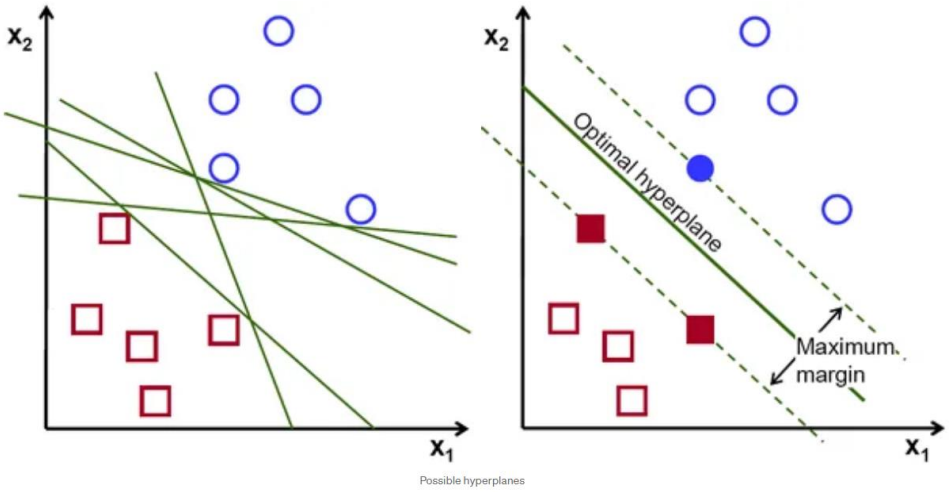

Figure 1. Hyperplanes and Support Vector.

In the scenario of a binary classification problem with two distinct classes, the dataset consists of input feature vectors labeled as X, each paired with their respective class labels. The equation for the hyperplane is given by:

$$w^T + b = 0 \tag{1}$$

Here, W serves as the normal vector to the hyperplane, signifying the direction that is perpendicular to the surface of the hyperplane. The element b within the equation represents the hyperplane's offset or its distance from the origin in the direction indicated by the normal vector W.

Moreover, the decision boundary of the SVM is determined using optimal multipliers coupled with the respective support vectors. Training samples where i is greater than 0 are designated as support vectors. The decision boundary's equation is consequently described by the subsequent formula:

$$w = \sum_{i \to m} \alpha i * ti * K(xi * x) + b \tag{2}$$

$$ti(w^T * xi - b) = 1 \leftrightarrow b = w^T * xi - ti \tag{3}$$

2.3.2. Decision tree.

As shown in figure 2, a decision tree resembles a tree diagram, with each inner node representing a feature, the branches indicating decision rules, and the terminal nodes signifying the algorithm's outcome. Its potency is acknowledged in the realm of algorithms.

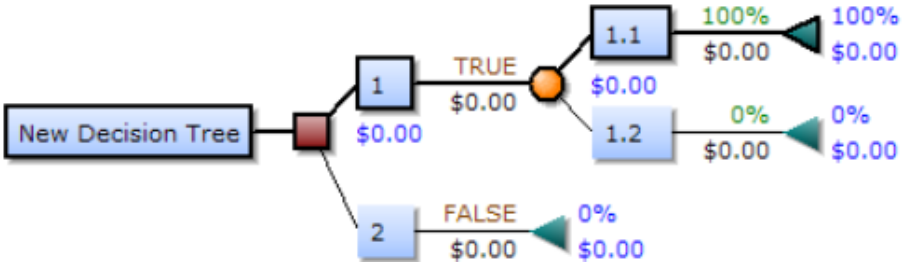

Figure 2. Decision tree building blocks [10].

2.3.3. Random Forest.

Random forest [11] is used for both regression and classification issues. using ensemble learning technique. The algorithm is constituted of several decision trees, and the ensuing 'forest' is educated using a technique termed as bagging or bootstrap aggregating. Bagging serves as a meta-algorithm within ensemble learning, enhancing the accuracy of various machine learning methods.

The training approach for random forest utilizes the method of bootstrap aggregating, commonly referred to as bagging, with tree-based learners. Given a training dataset X with associated responses Y, bagging involves repeatedly drawing a random subset from the training data (B times), and then building trees based on these subsets. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$f = \frac{1}{B} \sum_{b=1}^{B} fb(x') \tag{4}$$

Additionally, a way to estimate the prediction's uncertainty is by calculating the standard deviation of predictions generated by all individual regression trees for the input vector x':

$$\sigma = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (fb(x') - f)\text{^}2} \tag{5}$$

The optimal count of trees, B, is a parameter that can vary. Generally, anywhere from a few hundred to several thousand trees might be employed, contingent on the training data's characteristics and size.

3. Results

"STDs", "STDs number", "STDs Number of diagnosis", "STDs condylotomies", "Smokes years", "Dx Cancer" and "IUD" are variables that have higher correlation with other variables. Thus, these variables were removed. Consequently, the age of an individual's initial sexual encounter, the count of sexual partners, and the frequency of pregnancies, age, smoke years (packs), whether people are diagnosed as HPV and sexually transmitted diseases (STD) have all been linked to an increased risk of cervical cancer.

The figure 3 shows that women who aged between 20 and 40 are more likely to get cervical cancer. There is a noticeable depletion of the peak in each density plot looking at the age of the patients, demonstrating the significant association between "Age" and "Cervical Cancer".

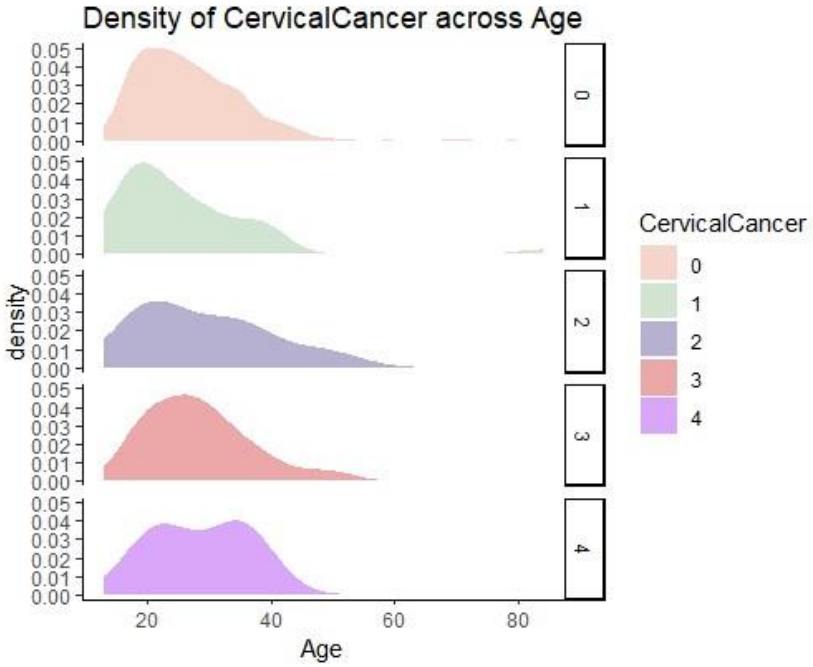

Figure 3. Density of Cervical Cancer Classes.

Additionally, there is a decreasing height of the peaks when one compares "Cervical Cancer" to "Hormonal contraceptives years" in the figure 4, and the right skewness of the density plot also implies a strong association.

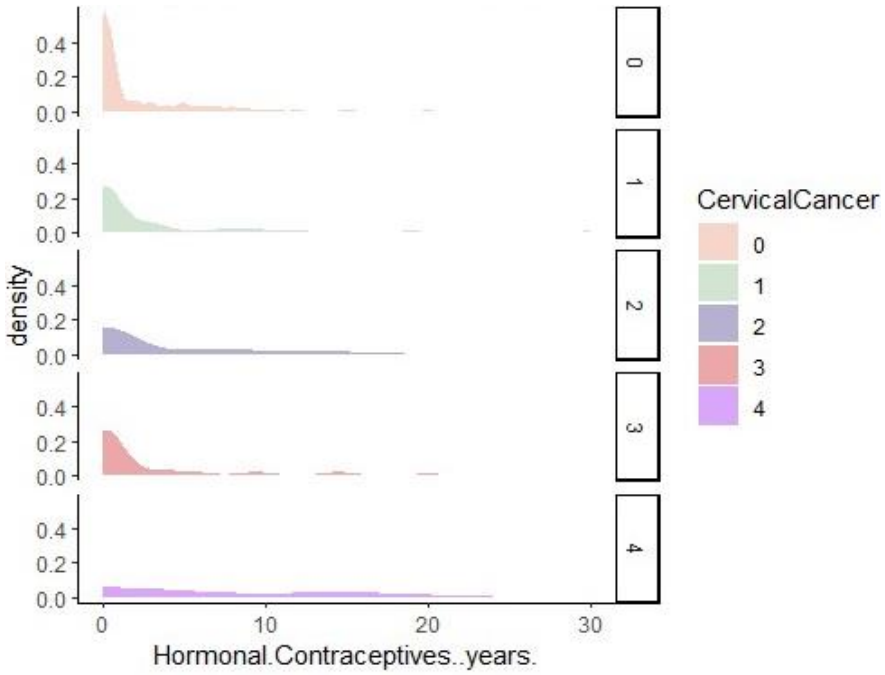

Figure 4. Density of Cervical Cancer across years of Hormonal contraceptives years.

After visualizing some of those independent variables, three feature selection methods were applied to this dataset, which are random forest, boruta, and lasso regression. The results of the three feature selection methods were compared separately with the results from visualization. Ultimately, boruta was chosen as the feature selection method. The final variables were selected, such as age, Dx, STD and smokes packs year.

By using SVM, decision tree and random forest, the result of prediction has been compared. All models were trained on a subset of the data (75%), and tested on the remaining 25%. The performance of the models was then evaluated using a confusion matrix. The accuracy, precision and recall were

used as metrics for the models' performance assessment. Precision (specificity) is about ensuring the model's accuracy, indicating the fraction of relevant examples among the retrieved examples. Recall (sensitivity), on the other hand, is about the model's ability to get the right instances.

Table 1. Assessment of those three models.

Model	Accuracy	95%CI	Sensitivity
SVM	0.8791	(0.8278, 0.9195)	0.8791
Decision Tree	0.8744	(0.8226, 0.9156)	0.8785
Random Forest	0.8744	(0.8226, 0.9156)	0.8821

The effectiveness of the models was assessed by evaluating their accuracy and recall. To ensure the model's performance, precision indicates the proportion of relevant instances among the retrieved examples. On the other hand, recall rate (sensitivity) refers to the model's capacity to identify the appropriate occurrences. The effectiveness of the model increases with higher values of precision and recall, both ranging from 0 to 1. Recall rate is more critical than precision in risky situations, such as identifying cervical cancer patients, because the cost of the incorrect treatment might be considerable, but not always. Finally, the random forest method was selected due to the fact that it has the highest recall rate (0.8821) and accuracy (0.8744) compared to the rest.

4. Discussion

Several research endeavors have sought to develop a reliable method capable of early-stage prediction and detection of cervical cancer using clinical data. In the study by Vidya et al., the dataset, consisting of five features, was partitioned into 500 samples for training and 100 samples for testing. When compared to alternative algorithms, it was observed that the Multilayer Perceptron exhibited the most favorable performance. MLP achieved an accuracy rate of 98%, a sensitivity rate of 98%, and an area under the ROC curve of 99%. In the study by using same dataset with this paper, Md Mamun Ali et al. analyzed the four factors separately in 2021. The Random Tree algorithm demonstrated the highest classification accuracy, achieving 98.33% for biopsy data and 98.65% for cytology data. Conversely, the Random Forest algorithm and the Instance-Based K-nearest neighbor algorithm delivered superior performance, yielding accuracy rates of 99.16% for Hinselmann data and 98.58% for Schiller data, respectively. In this research, a composite variable termed "cervical cancer" was created by integrating four factors. This analysis illustrates how predictions concerning cervical cancer can aid healthcare professionals in predicting the evolution of cervical cancer subsequent to its early-stage detection. This is accomplished by leveraging significant and commonly utilized clinical features. The effectiveness of the employed classifiers was evaluated through performance metrics, including accuracy and recall. Random Forest was ultimately chosen as the preferred model due to its good performance. Random Forest typically provides highly accurate predictions for cervical cancer, a critical aspect for early detection. By aggregating results from multiple decision trees, it reduces variance and enhances prediction stability and reliability. In cervical cancer research, there may be an imbalance between normal and cancer samples. Random Forest effectively addresses imbalanced data, improving the model's ability to predict cervical cancer.

Nevertheless, it should be noted that in practical applications, the available quantity of cervical cancer data was insufficient to comprehensively address these issues. Through further analysis of larger datasets, a more thorough understanding of the limitations inherent in this methodology was anticipated, particularly concerning the identification of cervical cancer and similar lesion types. Furthermore, enhanced precision may be achieved if data regarding the HPV infection status of cervical tissues can be incorporated. There remain several critical areas for future exploration. Firstly, it is crucial to assess the applicability of this study beyond the dataset sourced solely from UCI to other countries or regions. Secondly, while this study compares three machine learning algorithms for cervical cancer prediction, there is an array of alternative methods that warrant investigation. Lastly, this dataset contains variables with substantial amounts of missing data, which were not

addressed in this study. Consequently, gathering additional data for these variables may enhance the predictive accuracy of our model.

5. Conclusion

This study examines various risk factors associated with cervical cancer and employs three different approaches for the classification of a cervical cancer dataset. Ultimately, the random forest model was selected as the model for identifying cervical cancer cases. The study contributes valuable insights into the risk factors related to cervical cancer. Identifying these factors can potentially lead to higher rates of successful treatment and improved survival rates among female patients with cancer. Applying integration of machine learning significantly facilitates the prediction of appearance of cervical cancer.

References

[1] The overview of cervical cancer, *World Health Organization*.

[2] Naif Al Mudawi (2022). A Model for Predicting Cervical Cancer Using Machine Learning Algorithms, *MDPI, 22 (11), 4132*.

[3] Prabhpreet Kaur (2019). Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification, *ScienceDirect*.

[4] Asadi, F (2020). Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer. *J. Biomed. Phys. Eng. 2020, 10, 509 – 513.*

[5] Ching-Hsien Hsu (2021). Interactive personalized recommendation systems for human health, *Journal of Ambient Intelligence and Humanized Computing*.

[6] Anuraga G (2019). Random forest prognostic factor in colorectal cancer, *Journal of Physics: Conference Series*.

[7] Alyafeai Z (2020). A fully-automated deep learning pipeline for cervical cancer classification.

[8] Jundong Li (2017). Feature Selection: A Data Perspective. ACM Computing Surveys. 50, 6, Article 94, 45 pages.

[9] Rohith Gandhi (2018). Support Vector Machine — Introduction to Machine Learning Algorithms, *Towards Data Science*.

[10] abhishek (2023). Decision Tree, *GeeksforGeeks*.

[11] Onesmus Mbaabu (2020). Introduction to Random Forest in Machine Learning.