Cervical Cancer Risk Prediction and Analysis using Machine Learning Approach

Nathaniel Susianto Sutanto[a], Daniel Aditya Tumansery[a], Marcell Kurniawan Sutanto[a], Ahmad Husain[a]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Cervical cancer is a leading cause of death in females worldwide despite advances in prevention, diagnosis, and treatment. This study explores the use of machine learning techniques for cervical cancer risk prediction and analysis. With the advancements in computational power and availability of data, machine learning has a potential to help aid medical professionals in providing diagnoses, such as cervical cancer diagnosis. In this paper, machine learning was used to do risk prediction and analysis towards cervical cancer. As a target variable, Schiller test results was chosen out of the 4 available screening tests namely Hinselmann, Schiller, Biopsy and Citology. The data would go through oversampling and feature selection using Recursive Feature Elimination (RFE) for pre-processing. Finally, the risk prediction of cervical cancer can be obtained by classification using models such as SVM (Support Vector Machine), KNN (K-Nearest Neighbor), Random Forest and XGBoost. The result of this classification showed that XGBoost performed the best in predicting Schiller test result with an accuracy of almost 95%. A comparison of features used indicates that is due to our results showing that the accuracy scores did not change much when using only 6 features instead of the original 30 features. Some of the identified features or risk factors that could contribute to the diagnosis of cervical cancer are age, number of sexual partners, and age of first sexual intercourse.

Keywords: cervical cancer; features; risks; accuracy

1. Introduction

Cervical cancer is a type of cancer that develops in the cells of the cervix, which is the lower part of the uterus (womb) that connects to the vagina. It normally starts with abnormal alterations in the cervix's cells, known as precancerous cells, which can progress to cancer if left untreated. Cervical cancer is a leading cause of death in females worldwide despite advances in prevention, diagnosis, and treatment [1]. The most common cause of cervical cancer is long-lasting infection with human papillomaviruses (HPV). The majority of cervical malignancies are caused by chronic infections with particular forms of the sexually transmitted human papillomavirus (HPV) [2]. HPV is quite common, and almost all sexually active people will develop it at some point in their life. However, in the vast majority of cases, the immune system clears the infection without any long-term consequences. However, in certain situations, the infection can remain and eventually contribute to the development of cervical cancer.

Cervical cancer usually spreads slowly, allowing for early detection and treatment with routine screenings such as a Pap test (Pap smear), Schiller test, Hinselmann test, and cervical biopsy [3]. These tests can detect precancerous abnormalities in the cervix, enabling interventions to prevent cancer or find cancer at an early stage when it is most curable. Atypical vaginal bleeding (such as bleeding between periods, after sex, or after menopause), pelvic pain, pain during sexual intercourse, and atypical vaginal discharge are all symptoms of cervical cancer [4]. Cervical cancer prevention generally entails frequent screening, HPV vaccination, and safe sexual practices through the use of contraceptives. The HPV vaccine is advised for both males and females between the ages of 9 and 26 before they begin sexual activity. It protects against the most common HPV strains linked to cervical cancer.

Cervical cancer is a type of cancer that attacks a woman's intimate organs, specifically the cervical area, which is the lowest part of the uterus. To date [5], there have been hundreds of thousands of cases of cervical cancer and tens of thousands of deaths in women ranging in age from adolescence to adulthood. Cervical cancer is the world's second leading cause of cancer death among women. According to WHO [6] 500,000 women worldwide are diagnosed with cervical cancer each year, with an estimated 300,000 dying as a result. Cervical cancer fatality rates in women rise year after year. The death rate of women with cervical cancer is very high, with over 60% of them coming from the middle and lower middle classes, and the majority of them occurring in persons who do not have the money to do checks, early identification, and treatment of cervical cancer in their bodies.

Then the global cancer data [7] revealed that cervical cancer is the fourth most common disease among women, with a mortality rate of around 90% in underdeveloped and developing countries due to lack of public knowledge about the causes

and effects of cervical cancer. namely knowledge of HPV (Human Papilloma Virus). The main cause of cervical cancer is HPV (Human Papilloma Virus). This virus can harm cervical cells, squamous cells and gland cells. These precancerous cells are referred to as Cervical Intraepithelial Neoplasia (CIN), which affects the surface tissues of the cervix only. Over time, a small percentage of CIN will develop into cancer. For that it takes a period of two to three decades for cervical cancer to reach an aggressive state, early detection and proper treatment can significantly reduce this disease. Factors associated with the development of cervical cancer include sexual activity starting at a young age (less than 16 years), a high total number of sexual partners (more than four), and a history of genital warts. Actually, vaccines for viruses such as HPV 16 and HPV 18 are now available in the market [8], but due to lack of awareness, knowledge as well as good socialization, these preventive measures are not commonly used. Some simple precautions but have a big impact, namely awareness of the occurrence of cervical cancer, early marriage or at a very young age, prolonged use of contraceptive pills, cleanliness of intimate organs, many sex partners or exchanging partners, and immunity or body resistance. low. However, of these several actions, the most important thing that must be done to prevent cervical cancer is by carrying out early detection and proper treatment so that it can reduce the incidence and risk of death that is not detected significantly.

Machine learning, a subfield of artificial intelligence, has gained popularity in recent years with vast potential in various industries, including the field of diagnosis and treatment of diseases [9]. Machine learning has shown immense promise in predicting diagnostics, treatment planning, and patient outcomes. By utilizing computational power to learn from data and make accurate predictions, machine learning has the power to enhance medical research, simplify clinical workflows, and improve patient care. In the medical field, data available and gathered is continually expanding. Patient health records, medical diagnostics, genomics, and clinical trials produces a set of complex data that is often difficult to be processed and analyzed manually by humans [10]. This is where machine learning offers a solution, it could help harness the potential of this data and extract hidden insights through the utilization of data mining. By leveraging advanced algorithms and statistical models, machine learning algorithms can process vast datasets, identify patterns, detect anomalies, and generate information to support medical decision-making. one downside of using machine learning is that the computational model could potentially overfit the datas supplied to it. Feature selection is the process of selecting only relevant features to be used in training a machine learning model in order to prevent overfitting which could result in inaccuracy of the model's prediction.

One of the major contributions of machine learning in medicine is the ability to detect and diagnose potential diseases. Machine learning algorithms can learn from automatically from medical records and genetic information to recognize patterns and risk factors of diseases that would otherwise be missed by human observers. This can lead to an early detection of diseases that could ultimately save lives [11]. This is where machine learning approach could come in and help prevent early stages of cervical cancer in women. By analyzing clinical data collected from hospitals worldwide, a trend or risk factors can be found and could potentially help medical professionals in prioritizing prevention or treatment for the patient. The novelty of this research is in the utilization of feature selection to process the cervical cancer risk factor dataset to find which clinical data has the most influence to the results of various cervical cancer tests. This could help medical professionals to only collect the relevant data for patient screening and reduce the complexity of data collection and processing of future cervical cancer data.

2. Related Works

Numerous work has been done regarding the classification of cervical cancer risk, some of which has also utilized machine learning, and this dataset in particular. Research by Deng X. et al. utilizes three machine learning models namely SVM, XGBoost and random forest [12]. One interesting technique used is oversampling which creates synthetic data to equalize the class distribution of the target variable. The oversampling method used is borderline-SMOTE. Lu J. et al. combined the existing method of classification model training with genomic sequencing dataset [13]. Although the end result did not seem to improve much compared to the standalone machine learning model. Nithya B. and Ilango V. evaluated and compared multiple feature selection approaches on this cervical cancer dataset [14]. Feature selection models such as RFE and Boruta are compared to find the features that has the highest impact towards cervical cancer diagnosis. Their results found that using only 10 out of the 26 features available increases the accuracy of all models trained for the dataset therefore signifying the importance of feature selection to prevent overfitting.

Gupta A. et al utilized a recall-based approach in evaluating their random forest model's effectiveness in predicting cervical cancer with this same dataset [15]. They also used SMOTE to upsample the data and analyzed all four of the target variables. Overall, the obtained recall scores are very satisfactory with the highest recall achieving 0.996 when predicting biopsy results with only 1 case of false negative. Finally, a research by Sagala N. compared the performance of 3 classification models, namely SVM, Naive Bayes and KNN [16]. The results showed that reducing number of features using random forest and correlation-based filter approaches increases accuracy for both naive Bayes and KNN models. Meanwhile accuracy for SVM experienced a decrease despite using the most computational power compared to the other two models.

Another look at performed with the aid of using Wu and Zhou [17] carried out 3 exceptional techniques to diagnose 4 goal variables of cervical cancer: HSCB. The dataset contained 32 elements that probably reason cervical cancer. The authors used the Random Oversampling technique to stability the dataset of 668 patients. They blended SVM with Recursive Feature Elimination (RFE), blended SVM with Principal Component Analysis (PCA), and hired the conventional SVM, and that they as compared those technique's characterization of cervical cancer. The class consequences display that SVM–RFE and SVM–PCA offer greater correct consequences whilst deciding on best 8 elements in preference to the use of the conventional SVM techniques [18].

Using RFE and PCA to exclude many variables, Abdoh et al. [18] predicted cervical cancer. To balance the UCI cervical cancer dataset, the researchers employed the Synthetic Minority Oversampling Technique (SMOTE). The Random Forest (RF) method was used to diagnose cervical cancer, and it was compared to the RF-PCA and RF-RFE methods. When compared to the work of Wu and Zhou [17], their results suggest that incorporating SMOTE with RF classifiers improves classification accuracy by roughly 4%.

Adem et al. [19] employed Deep Learning to predict cervical cancer using Softmax classification with a stacked autoencoder on the same dataset. The softmax layer was utilized to predict HSCB of cervical cancer, while the stacked autoencoder was used as a dimensional reduction method. The authors used six ML approaches to test their approach: RF, Decision Tree (DT), MLP, SVM, Rotation Forest models, and KNN [19]. Their proposed method for diagnosing the four target variables (HSCB) of cervical cancer performed better than Wu and Zhou's method [17], with a 4% improvement in classification accuracy.

3. Methodology

In this section, we present the methodology of our study in detail. Our aim is to predict indicators/diagnosis of cervical cancer using several machine learning approaches. We will provide detailed explanations of each technique and algorithm in the following sections. Below is a visualization of the methodology workflow "Figure. 1" which this research used.

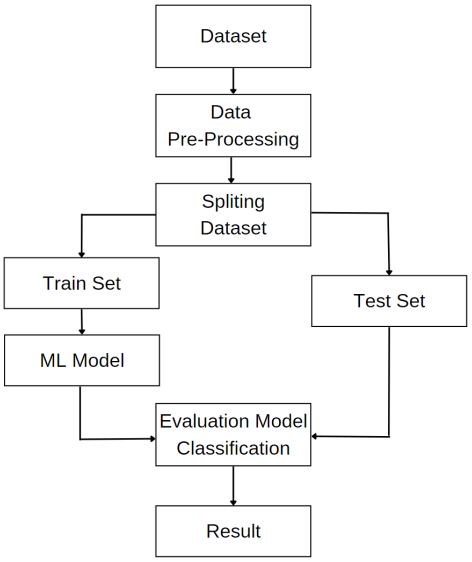

Figure 1. Methodology workflow

3.1 Dataset

The dataset used is sourced from UCI Machine Learning Repository and can be accessed through the following link https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. UCI Machine Learning Repository is a collection of databases that could be used to explore various machine learning algorithms, the repository itself has been cited numerously by various research publications. The data itself originated from 'Hospital Universitario de Caracas', a university hospital in Caracas, Venezuela. A total of 858 patient data was recorded for 36 data attributes, 4 of which are the labels for various cervical cancer detection methods as can be seen in Table 1. It needs to be noted that the positivity of each target variable does not guarantee cervical cancer, instead, it only signifies a high risk of having one.

Field Name	Type (Format)	Field Name	Type (Format)

Age	Integer	**STDs:pelvic inflammatory disease**	Boolean
Number of sexual partners	Integer	**STDs:genital herpes**	Boolean
First sexual intercourse	Integer	**STDs:molluscum contagiosum**	Boolean
Num of pregnancies	Integer	**STDs:AIDS**	Boolean
Smokes	Boolean	**STDs:HIV**	Boolean
Smokes (years)	Integer	**STDs:Hepatitis B**	Boolean
Smokes (packs/year)	Integer	**STDs:HPV**	Boolean
Hormonal Contraceptives	Boolean	**STDs: Number of diagnosis**	Integer
Hormonal Contraceptives (years)	Integer	**STDs: Time since first diagnosis**	Integer
IUD	Boolean	**STDs: Time since last diagnosis**	Integer
IUD (years)	Integer	**Dx:Cancer**	Boolean
STDs	Boolean	**Dx:CIN**	Boolean
STDs (number)	Integer	**Dx:HPV**	Boolean
STDs:condylomatosis	Boolean	**Dx**	Boolean
STDs:cervical condylomatosis	Boolean	**Hinselmann (Target)**	Boolean
STDs:vaginal condylomatosis	Boolean	**Schiller (Target)**	Boolean
STDs:vulvo-perineal condylomatosis	Boolean	**Citology (Target)**	Boolean
STDs:syphilis	Boolean	**Biopsy (Target)**	Boolean

Table 1. Dataset label description

3.2 Data preprocessing and feature selection

Of the 4 methods, the ‘Schiller’ method was chosen as the main target variable. From the remaining 32 features, 2 of which are dropped, namely ‘STDs: Time since first diagnosis’ and ‘STDs: Time since last diagnosis’. This is due to the numerous null values it contains therefore nullifying any relevancy towards the final target prediction. Other null values were also found in the remaining columns, these null values are then replaced with median of the column’s values. Further logical checking are also done to ensure data is logical. Such as age must be above age of first intercourse, years of IUD and smoking must also be smaller than age.

The data will also be upsampled due to the imbalance of the 0 and 1 class from the ‘Schiller’ column as seen in figure 2, therefore an oversampling algorithm will be employed. Oversampling is a technique where synthetic data will be generated for the minority class. The oversampling will be done using SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) algorithm [19], it is essentially an extension of the SMOTE algorithm that is suited for both categorical and continuous data like the ones in the cervical cancer dataset.

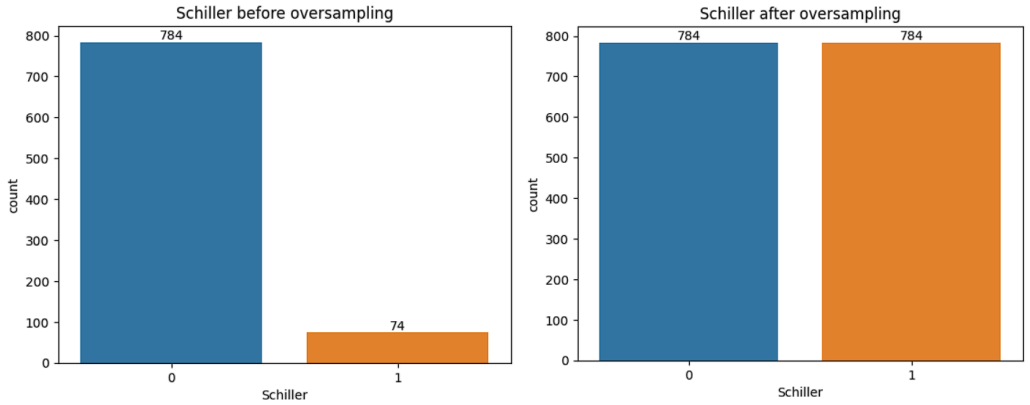

Figure 2. Data before (left) and after oversampling (right)

Finally, the data will go through feature selection using RFE (Recursive Feature Elimination). RFE is a feature selection algorithm that recursively eliminates unimportant features until eventually the number of features left are only as desired. In this case, 6 features are specified as the target feature count or only 20% of the total features, this is because the dataset itself contains a lot of boolean values which are not relevant to classification. Booleans such as smokes, IUD, etc. are not necessary since if the value is 0, then the years will be more than 0. The RFE model was fit with a decision tree classifier that is simple enough to implement yet it provides a good classification score when the data is trained.

After data has been preprocessed, a machine learning model will be trained on said data. For this research, multiple classification models will be trained on the data to see how it performs on classifying cervical cancer risks. The classification models used are SVM (Support Vector Machine), KNN (K-Nearest Neighbor), Random Forest, and XGBoost. In total, two set of classification models (8 models in total) will be trained, one set with only 6 features chosen by RFE, while the other with all 30 features. These models will be evaluated using 20% of each respective dataset while 80% of it will go to training of the model. Model will be evaluated and compared using test accuracy and precision. In the medical field, especially diagnosis, it is very important to evaluate precision scores in order to reduce the case of false negatives which could lead to problems when the model actually gets deployed to the real world scenario.

4. Result & Analysis

4.1 Classification Results

The data will undergo Recursive Feature Selection in python to reduce the 30 features to only 6 important ones that are the most correlated with the Schiller test results of the patients. The classification models are then trained individually with the data. Parameters for certain algorithm such as SVM, KNN and Random Forest has been determined by Grid Search to ensure the best possible outcome. Table 2 shows the resulting accuracy and recall of each model after being trained and tested on an 80-20 split dataset. From the table, there is not a lot of variations between algorithms with most of it achieving a relatively good accuracy score above 0.8. Random forest and XGBoost in particular performs very well with an accuracy almost reaching 95%. In general, every algorithm obtained a lower accuracy when using 6 features instead of 30, but the decrease is quite negligible for most algorithm except for KNN that experienced around a 3% decrease in both accuracy and recall. As for recall, the two best performing algorithm gained an increase in recall score when RFE is applied. This means applying RFE has a slightly positive impact on predicting overall true values.

	acc (30 features)	**acc (6 features)**	**recall (30 features)**	**recall (6 features)**
SVM	0.8854	0.8758	0.9051	0.8101
KNN	0.8662	0.8376	0.9241	0.8924
RandomForestClassifier	0.9490	0.9427	0.9114	0.9177
XGBoost	0.9554	0.9459	0.9241	0.9304
mean	0.9140	0.9005	0.9161	0.8877

Table 2. Accuracy and Recall of each model trained on RFE and original data

These accuracy and recall scores showed that there are a lot of irrelevant features in the dataset when it comes to predicting the result of the Schiller test. Out of the 30 predictive features, 6 of them were sufficient in predicting the results, with only a slight decrease in accuracy score. The best performing model is XGBoost with both the best accuracy and recall score. Random forest also managed to achieve a good accuracy trailing behind XGBoost. Prediction results of XGBoost can be seen in Figure 3 in the form of its confusion matrix. As seen in Figure 3, XGBoost performs excellently when predicting false labels, but in predicting true labels, it still fails to catch some cases.

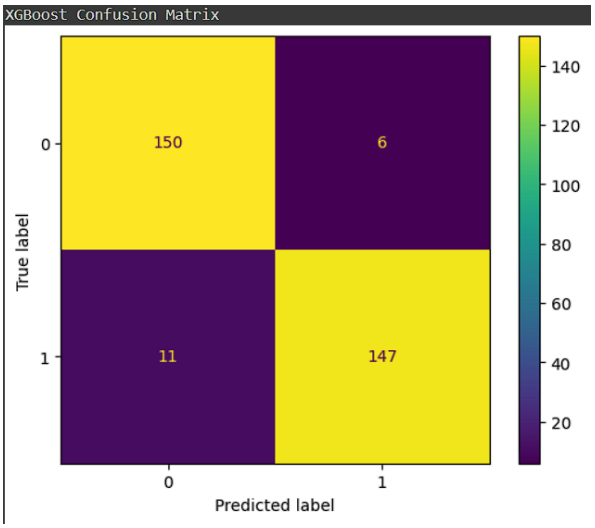

Figure 3. Confusion matrix of XGBoost classification

4.2 Important Features

The 6 features that have been identified as having high importance on predicting Schiller test results can be found in Figure 4. Most of the 6 used features are continuous numerical values instead of boolean, except for 'Hormonal Contraceptives' which is a boolean that determines if the patient has ever used hormonal contraception. All 6 of the variables are theoretically correlated with cervical cancer risk, age can be a huge factor in developing cervical cancer, as a women gets older, their chances of developing this disease increases as well. Early sexual activity could also be a risk factor in developing cervical cancer, this could be represented by the data 'Number of sexual partners', 'First sexual intercourse' and 'Num of pregnancies'. Hormonal contraceptives could also possibly cause this type of cancer since it messes with thebody's natural hormonal cycles that could cause abnormalities in the reproductive system cells. But one of the most important risks of cervical cancer is surprisingly not considered as an important feature. HPV is the greatest risk factor that could determine cervical cancer, therefore it is expected to make it into one of the 6 risk factors. HPV is possibly not included to the features list due to it being a boolean value and the model having a slight bias towards boolean value where not a single categorical boolean is picked.

	Risk Factors/Features
1	Age
2	Number of sexual partners
3	First sexual intercourse
4	Num of pregnancies
5	Hormonal Contraceptives (years)
6	Hormonal Contraceptives

Table 3. Features after RFE

5. Conclusion

In conclusion, the machine learning models trained on classifying the UCI cervical cancer dataset has successfully classified cervical cancer, specifically, using data of the patient's Schiller test. All of the used machine learning models, only experienced a slight decrease in accuracy when the features are reduced from its original 30 features to only 6 features. The model with best accuracy and recall is XGBoost model with an accuracy of around 0.95 when predicting the oversampled data using only 6 features. Further analysis into the confusion matrix, shows that this model could very accurately predict negative cases of cervical cancer, while its prediction on positive cases could be improved further.

Recursive Feature Elimination also found that the 6 features that are the most relevant in predicting Schiller's test result is age, number of sexual partners, age of first sexual intercourse, number of pregnancies, years of using hormonal contraceptives and usage of hormonal contraceptive. This is inline with some known risk factors of cervical cancer such as sexual activity and age. But one STD in particular, which is HPV did not make much of a threat when it comes to the outcome of Schiller's test. This could be due to oversampling of the data that causes an error in identifying HPV as an important risk factor. Or it could also be that the current dataset just do not contain a lot of HPV cases to justify using HPV as an important risk factor. As a suggestion to further research, it could be a good idea to combine this dataset with some other available ones in order to equalize the class of diagnosis. Some other feature selection methods could also be used in determining risk factors such as PCA, Boruta or ADASYN.

Reference

[1] S. Zhang, H. Xu, L. Zhang, and Y. Qiao, "Cervical cancer: Epidemiology, risk factors and screening," *Chinese Journal of Cancer Research*, vol. 32, no. 6, pp. 720–728, 2020. doi:10.21147/j.issn.1000-9604.2020.06.05

[2] N. Kashyap, N. Krishnan, S. Kaur, and S. Ghai, "Risk factors of cervical cancer: A case-control study," *Asia-Pacific Journal of Oncology Nursing*, vol. 6, no. 3, pp. 308–314, 2019. doi:10.4103/apjon.apjon_73_18

[3] P. Cafforio *et al.*, "Liquid biopsy in cervical cancer: Hopes and pitfalls," *Cancers*, vol. 13, no. 16, p. 3968, 2021. doi:10.3390/cancers13163968

[4] C. A. Burmeister *et al.*, "Cervical cancer therapies: Current challenges and future perspectives," *Tumour Virus Research*, vol. 13, p. 200238, 2022. doi:10.1016/j.tvr.2022.200238

[5] Andrian, Steele, E. S. Salim, Hartato Bindan, Endy Pranoto, and Abdi Dharma, "Analisa Metode Random Forest Tree dan K-Nearest Neighbor dalam Mendeteksi Kanker Serviks", *JIKOMSI*, vol. 3, no. 2, pp. 97-101, Sep. 2020.

[6] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and Prognosis," *Cancer Informatics*, vol. 2, p. 117693510600200, 2006. doi:10.1177/117693510600200030

[7] H. Lin, Y. Hu, S. Chen, J. Yao, and L. Zhang, "Fine-grained classification of cervical cells using morphological and appearance based Convolutional Neural Networks," *IEEE Access*, vol. 7, pp. 71541–71549, 2019. doi:10.1109/access.2019.2919390

[8] K. Hemalatha and K. U. Rani, "An optimal neural network classifier for Cervical Pap smear data," *2017 IEEE 7th International Advance Computing Conference (IACC)*, 2017. doi:10.1109/iacc.2017.0036

[9] K. Arun Bhavsar *et al.*, "Medical Diagnosis Using Machine Learning: A statistical review," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 107–125, 2021. doi:10.32604/cmc.2021.014604

[10] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big Data in Healthcare: Management, analysis and future prospects," *Journal of Big Data*, vol. 6, no. 1, 2019. doi:10.1186/s40537-019-0217-0

[11] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *Healthcare*, vol. 10, no. 3, p. 541, 2022. doi:10.3390/healthcare10030541

[12] X. Deng, Y. Luo, and C. Wang, "Analysis of risk factors for cervical cancer based on machine learning methods," *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2018. doi:10.1109/ccis.2018.8691126

[13] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," *Future Generation Computer Systems*, vol. 106, pp. 199–205, 2020. doi:10.1016/j.future.2019.12.033

[14] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," *SN Applied Sciences*, vol. 1, no. 6, 2019. doi:10.1007/s42452-019-0645-7

[15] A. Gupta, A. Anand, and Y. Hasija, "Recall-based machine learning approach for early detection of cervical cancer," *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021. doi:10.1109/i2ct51068.2021.9418099

[16] N. T. Sagala, "A comparative study of data mining methods to diagnose cervical cancer," *Journal of Physics: Conference Series*, vol. 1255, no. 1, p. 012022, 2019. doi:10.1088/1742-6596/1255/1/012022

[17] W. Wu and H. Zhou, “Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches,” IEEE Access, vol. 5, pp. 25189–25195, 2017, doi: 10.1109/access.2017.2763984.

[18] S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, “Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques,” IEEE Access, vol. 6, pp. 59475–59485, 2018, doi: 10.1109/access.2018.2874063.

[19] K. Adem, S. Kiliçarslan, and O. Cömert, “Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification,” Expert Systems with Applications, vol. 115, pp. 557–564, Jan. 2019, doi: 10.1016/j.eswa.2018.08.050.

[20] N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research, 321-357, 2002.