From Coverage to Costs: Multi-Model Analysis of Factors Shaping Cervical Cancer Prevention

Xinyang Qian *

Nanjing Foreign Language School, Nanjing, 210000, China

* Corresponding Author Email: qian.xinyang@outlook.com

Abstract. World Health Organization (WHO) is developing a global strategy for cervical cancer prevention to scale the human papillomavirus vaccination coverage to 90%. To measure the weight of predictors affecting cervical cancer prevention effect, this paper analyzes the contribution of factors in both statistical and machine learning methods, including logistics regression, multinominal logit regression, random forest, extremely randomized forest, GBDT, and XGBoost. Data processing and model effectiveness analysis comparison are done, varying vaccination coverage, cost, region, and assumptions. The paper finds that current cost, current mortality prevention, projected cost, projected cost prevention, and current HPV vaccination coverage are leading factors influencing future cervical cancer prevention. In contrast, the current CC prevention rate is the major factor indicating current HPV vaccination coverage. Predictions are consistent across the six models. In conclusion, although high vaccination coverage can reduce CC infection rates, the high cost of vaccination still affects vaccine coverage and thus the effectiveness of cervical cancer prevention.

Keywords: Cervical cancer; vaccination coverage; projected cost; logistics regression; decision tree.

1. Introduction

Human papillomavirus (HPV) infection unequivocally stands as a primary causative factor in the development of cervical cancer [1]. WHO is developing a global strategy towards eliminating cervical cancer as a public health problem, which proposes an elimination threshold of four cases per 100000 women and includes 2030 triple-intervention coverage targets for scale-up of human papillomavirus (HPV) vaccination to 90% [2]. The ultimate goal of HPV vaccination is the eradication of high-risk HPVs [3]. High-risk HPV types, such as HPV16, HPV18, HPV31, HPV33, and HPV45, are associated with a higher risk of developing cervical cancer [4].

Against most high-risk HPV types, gender-neutral vaccination has a strong impact already when moderate cross-protection and vaccination coverage apply, according to the prevalence of HPV18/31/33/35 in 2011-2014 by birth cohort (1992-1995) among non-vaccinated 18-year-old Finnish girls following earlier gender-neutral vaccination of the same birth-cohorts in 2007-2009 [5]. HPV infection is typically detected in early-stage cervical lesions, such as low-grade squamous intraepithelial lesions and high-grade squamous intraepithelial lesions [1]. As the disease progresses, HPV integration becomes more prevalent, acting as a potential carcinogenic factor. Despite being a curable cancer if detected early, cervical cancer contributed to 9.4 percent of new cancer cases and 70,000 deaths in India in 2020 [6].

Justifying the prediction of previous research and accumulating evidence from case studies in both Sweden, a high-income country, and India, a developing country, the author accessed baseline results of enrolling for HPV vaccination, the results of the HPV genotyping tests obtained until 2022-12-31 and the 3-year model predictions on HPV incidence concluded by a population-based trial offering all Swedish women (aged 23–29) concomitant HPV vaccination and HPV-screening [5]. Informed with the crude incident rates of HPV-related cancers, burden of cervical cancer, burden of HPV infection in India, and comparison of the ten most frequent HPV oncogenic types in India among women with and without cervical lesions, the author accessed India: Human Papillomavirus and Related Cancers, Fact Sheet 2023 from ICO/IARC Information Centre on HPV and Cancer [7]. Therefore, the author realized that the region, income group, or other socioeconomic factors can be the potential predictors of the cervical cancer prevention effect.

Here, this paper examines the clinical and demographic datasets from cohorts of cervical cancer patients belonging to different races and regions, including 176 countries, encompassing Europe and Central Asia (28%), Sub-Saharian Africa (26%), and Others (46%): Cervical Cancer Vaccines [8]. The paper obtains overlapping and distinct relationships by organizing and analyzing data through various statistics and machine learning models, claiming that the HPV vaccination coverage and the cost invested in cervical cancer prevention are also predictors. Based on the novel discovery and the population-based reports accessed, this paper is going to compare the weight of both categorical variables describing socioeconomic factors and the quantitative variables measuring public health factors, while applying an innovative approach based on the state-of-the-art data analysis.

Recently, machine and deep learning models have been developed and utilized to enhance research efforts in cervical cancer detection through Pap smear images, thus experiencing progress in cervical cancer prevention. In a research study conducted by Kalbhor et al., the discrete cosine transform (DCT) and discrete wavelet transform (DWT) were employed to extract features, which provides input for seven machine learning classifiers to differentiate between various subgroups of cervical cancer, achieving an accuracy of 81.11% [9, 10]. Moreover, machine learning (ML) and Deep learning (DL)-based computer-aided diagnostic (CAD) techniques are being extensively expanded to automatically segment and categorize cervical cytology and colposcopy images [11]. This series of related research results inspires the author to use innovative methods in related fields to solve problems that urgently need to be solved in real life.

In summary, cervical cancer prevention reflects the Humanitarian care for women’s health, regional differences in health conditions, and alleviation of population mortality, as well as inspecting the essence behind the phenomenon through the nuances between the results obtained by different learning preferences and methods.

2. Methods

2.1. Data Source

Clinical and demographic data sets from cohorts of cervical cancer patients belonging to different races and regions on the Kaggle website were used in this study: Cervical Cancer Vaccines. This paper chooses the data in 2021, excluding constant data such as future HPV vaccination coverage (0.9) according to the goal set by WHO, and excluding data combinations with excessive correlation and collinearity.

2.2. Variable Selection

The data used in this paper counts 176 countries and 14 variables, including region, income group, cohort size, current HPV vaccination coverage, current HPV vaccination cohort size, future HPV vaccination cohort size, current cervical cancer prevention, current mortality prevention, projected cervical cancer prevention, projected mortality prevention, current cost, current cost prevention, projected cost, and projected cost prevention.

2.3. Data Processing

This paper aims to discuss the current impact of HPV vaccination coverage on cervical cancer prevention. Initially working with solely numerical variables, the author introduced new categorical variables in order to apply Ordered Logistic Regression (OLR), leading to a holistic interpretation. By computing the ratio of prevented cases to the vaccination cohort, dividing the numerical values into different ranges, recording those data as “curr/future_vacc_cc_prev”, and converting the data into percentage form to amplify and visualize data nuances and variations to reveal more significant correlations and contrasting effects, the author classified the cervical cancer prevention into categories (very low: 0~25 percentile, low: 25~50 percentile, high: 50~75 percentile, very high: 75~100 percentile) based on the magnitude of the “future_vacc_cc_prev percent”. The categories for cervical cancer prevention (very low, low, high, very high) have a clear order from low to high prevention, which is suitable for using the Ordinal Logistic Regression. To label each data point with a category, the author applied Excel to sort the data, added categorical variables, and imported the processed data into the SPSSAU online platform to generate OLR analysis results.

2.4. Statistical Methods

Starting with ordered logit to identify general trends, the author then used Multinominal Logit Regression to explore specific differences in order to model the relative probabilities of being in one of the categories compared to a reference category, allowing for non-constant odds across categories. These models allow the effects of independent variables to vary more flexibly across different categories, according to the Econometrics of Health and Medical Outcomes. Aware of the importance of the Collinearity Analysis in Multinominal Logit Regression, the author employed a correlation matrix to exclude redundant variants with correlation coefficients above 0.8, ensuring meaningful and interpretable results, enhancing the numerical stability of the regression model, and improving the accuracy of the estimated relationships between predictors and outcomes.

2.5. Machine Learning Methods

Initially analyzing with categorical variable cervical cancer prevention effect category, the analysis loses some accuracy by classifying data points within a certain interval. In order to achieve a rigorous conclusion, the author utilizes Random Forest and Extremely Randomized Trees in machine learning, dissecting the weight of different factors affecting the cervical cancer prevention effect and judging the optimality of the two ensemble learning algorithms based on decision trees. By cross-validation, the author obtained two sets of results, in order to either handle high-dimensional data sets with noisy data or deal with data sets with clear relationships and are less prone to overfitting, combining the advantages of both methods. Based on the previous methods, the author utilizes Gradient Boosting Decision Tree (GBDT) and eXtreme Gradient Boosting (XGBoost), employing multiple weak classifiers to perform iterative weighting and revising the fitting.

3. Results and Discussion

3.1. Ordinal Logistics Regression

Firstly, the author applied the Ordinal Logistic Regression to gain a comprehensive analysis of various possible factors that may affect cervical cancer prevention, thus identifying the significant, moderate, and minor factors. To investigate the factors with strong correlation, the paper focused on the major factors with a p-value<0.01, moderate factors with a p-value<0.05, and minor factors with a p-value<0.1, while maintaining the goodness of fit by ensuring that the McFadden R^2 is relatively large (Table 1). The R-squared computed through three distinct methods ensured the author to filter for models explaining a higher percentage of the variation in ‘cc_prev_category.’

The key inspection is that the current HPV vaccination coverage significantly affects the cervical cancer prevention category, by noticing the p-value=0.018<0.05. Besides, the mortality preservation category is closely tied to the cervical cancer prevention rate (p-value<0.01), which corresponds to its low survival rates, leading to its high mortality rate. While many researchers claim that cervical cancer remains a pressing issue, especially for low- and middle-income countries, the chart reveals that the income group is not a significant factor influencing cervical cancer prevention (p-value>0.1). This is probably because some wealthy countries lack awareness of the prevention of gynecological diseases and the strategic distribution of their resources, whereas the reverence and cherishment of life have been engraved in the cultural traditions and instincts of some relatively poor countries. By contrast, region acts as a moderate contributor to the cervical cancer prevention category (p-value=0.1), since infectious diseases are usually regional-based, depending on not only governmental and social efforts, but also the possibility for the HPV to spread and create a population-level effect.

Moreover, the current cervical cancer and mortality prevention are slightly more related to the cervical cancer prevention in the long term than the projected cervical cancer and mortality prevention.

Thus, prioritizing broad coverage early in vaccination campaigns can maximize future prevention efforts, reducing cervical cancer cases more effectively than solely focusing on maintaining uniform rates across populations over time. The prediction accuracy of the OLR is as high as 70.455%, with 83.333% for very high CC prevention, 79.545% for very low CC prevention, 73.333% for high CC prevention, and 46.667% for low CC prevention, highlighting the effectiveness of the model.

Table 1. Ordinal Logistics Regression.

item	item	regression coefficient	standard error	z -value	Wald $\chi 2$	p value	OR value
dependent variable threshold	high	2.570	0.712	3.607	13.012	0.000	0.077
	low	4.524	0.786	5.757	33.148	0.000	0.011
	very high	6.417	0.864	7.430	55.198	0.000	0.002
independent variable	current_cov	-1.474	0.621	-2.372	5.627	0.018	0.229
	proj_mort_prev	0.001	0.000	1.190	1.415	0.234	1.001
	region	0.128	0.078	1.645	2.705	0.100	1.137
	proj_cost	0.000	0.000	0.787	0.620	0.431	1.000
	income_group	-0.061	0.154	-0.394	0.155	0.694	0.941
	curr_mort_prev	-0.001	0.001	-1.328	1.764	0.184	0.999
	curr_cc_prev	0.001	0.001	1.543	2.381	0.123	1.001
	proj_cc_prev	-0.001	0.000	-1.525	2.327	0.127	0.999
	future_vacc_cohort_size	0.000	0.000	1.572	2.472	0.116	1.000
	mort_prev_category	1.783	0.202	8.846	78.257	0.000	5.950

3.2. Multinominal Logit Regression

Secondly, the author switched to the Multinominal Logit Regression to create a framework, offering a nuanced way to understand how different factors interact in various contexts of cervical cancer prevention (Table 2, 3 and 4). Generally, projective cervical cancer prevention is naturally correlated with the cervical cancer prevention effect ($p<0.05$), and the projective cost directly emphasizes the significance of investment in public health in the long term ($p<0.01$). Specifically, the author drew different conclusions for each category using the high CC prevention effect as a reference.

In regions with very low CC prevention effect (0~25 percentile), the income group doesn't seem to be a determining factor in the prevention outcome ($p>0.1$) (Table 2). Instead, what matters the most is the willingness to invest in HPV vaccination either currently or projected ($p<0.1$). This suggests that the barrier to improving prevention in these areas is not economic capability but the priority of health investments.

Table 2. Multinominal Logit Regression of Very Low.

very low	regression coefficient	standard error	z value	Wald $\chi 2$	p value	OR value	OR value 95% CI
proj_cc_prev	-0.001	0.000	-2.953	8.718	0.003	0.999	0.999 ~ 1.000
proj_cost	0.000	0.000	2.921	8.530	0.003	1.000	1.000 ~ 1.000
current_cov	2.661	2.234	1.191	1.419	0.234	14.313	0.179 ~ 1141.266
curr_cc_prev	-0.194	0.100	-1.941	3.769	0.052	0.824	0.678 ~ 1.002
curr_cost	0.000	0.000	1.765	3.115	0.078	1.000	1.000 ~ 1.000
region	0.154	0.140	1.102	1.213	0.271	1.166	0.887 ~ 1.534
income_group	0.056	0.258	0.216	0.047	0.829	1.057	0.638 ~ 1.751
intercept	-0.864	0.955	-0.904	0.817	0.366	0.422	0.065 ~ 2.741

In areas where CC prevention effect is very high (75~100 percentile), both income group and region influence the outcomes ($p<0.05$) (Table 3). For example, wealthier countries and certain regions with established healthcare systems and infrastructure tend to have better outcomes in terms of CC prevention. Moreover, cost continues to be a key factor. For instance, vaccination programs in these regions may be well-funded, but the costs of maintaining or expanding these efforts remain significant, influencing the long-term success of prevention strategies. This strengthens continuous investment while maintaining the well-being gifted by good basic conditions. However, current HPV vaccination coverage is not a significant predictor, plausibly because the initial vaccination rates of these countries are sufficiently effective with little internal variation.

Table 3. Multinominal Logit Regression of Very High.

very high	regression coefficient	standard error	z value	Wald χ2	p value	OR value	OR value 95% CI
proj_cc_prev	0.000	0.000	2.573	6.621	0.010	1.000	1.000 ~ 1.001
proj_cost	-0.000	0.000	-2.675	7.157	0.007	1.000	1.000 ~ 1.000
current_cov	-0.599	1.267	-0.473	0.224	0.636	0.549	0.046 ~ 6.580
curr_cc_prev	-0.001	0.000	-1.946	3.788	0.052	0.999	0.999 ~ 1.000
curr_cost	0.000	0.000	1.865	3.477	0.062	1.000	1.000 ~ 1.000
region	0.304	0.122	2.492	6.211	0.013	1.355	1.067 ~ 1.720
income_group	0.742	0.325	2.280	5.199	0.023	2.100	1.110 ~ 3.972

In areas where CC prevention effect is low (25~50 percentile), HPV vaccination coverage is relatively more significant than the other two cases (Table 2). For those countries that undertake some financial pressure but are able to afford vaccination on a certain scale, focusing on increasing vaccine uptake can significantly boost prevention efforts. This suggests that improving vaccination coverage should be the priority in these regions, as it holds great potential for enhancing cancer prevention rates.

Table 4. Multinominal Logit Regression of Low

low	regression coefficient	standard error	z value	Wald χ2	p value	OR value	OR value 95% CI
proj_cc_prev	-0.000	0.000	-2.425	5.880	0.015	1.000	0.999 ~ 1.000
proj_cost	0.000	0.000	2.835	8.036	0.005	1.000	1.000 ~ 1.000
current_cov	0.975	1.021	0.954	0.911	0.340	2.650	0.358 ~ 19.616
curr_cc_prev	-0.000	0.001	-0.375	0.141	0.708	1.000	0.999 ~ 1.001
curr_cost	-0.000	0.000	-0.077	0.006	0.939	1.000	1.000 ~ 1.000
region	0.029	0.132	0.219	0.048	0.827	1.029	0.794 ~ 1.334
income_group	0.228	0.221	1.033	1.067	0.302	1.257	0.815 ~ 1.938
intercept	-1.263	0.875	-1.443	2.083	0.149	0.283	0.051 ~ 1.572

3.3. Machine Learning Results

Thirdly, while the categorization of cervical cancer prevention effect into four levels is reasonable, the author shifted to incorporate Random Forest methods to analyze the numerical variables instead of categorical variables, comparing the nuances between the weight of each predictor contributing to the current CC prevention effect percentage and the future CC prevention affect percentage respectively. Currently, HPV vaccination coverage accounts for approximately 0.66 of the current cervical prevention effect observed globally. On the contrary, for the future CC prevention effect is influenced by a complex interplay of factors, ranging from the projected costs and projected mortality prevention to projected cost prevention and current vaccination coverage (Figure 1).

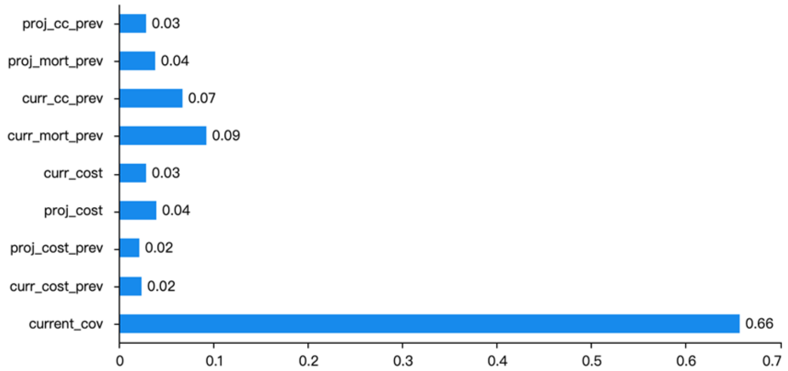

Fig 1. Random Forests (current CC prevention effect).

Primarily, Future projected costs of HPV vaccination and prevention programs will likely hold great weight on future cervical cancer prevention (0.28), since projections can account for factors like vaccine price inflation, increased healthcare infrastructure, and expanded access. Furthermore, projected prevention outcomes are closely linked to these future costs, as the affordability of widespread vaccination programs will determine their long-term success.

Secondarily, projected mortality prevention and projected cost prevention are moderate factors, both holding a weight of 0.16. Despite cervical cancer prevention naturally implies less projected mortality, since current mortality prevention rates are often impacted by the stage at which the cervical cancer is diagnosed and treated, mortality can be reduced through early detection and treatment programs, which are not considered in the model. Moreover, projected long-term cost savings from reduced cervical cancer cases make vaccination programs a cost-effective solution for cervical cancer prevention. While the initial cost of vaccination can be high, the reduction in treatment costs for cervical cancer cases can offset these investments in the long run.

Furthermore, while projected and current prevention of HPV infections is crucial to the goal of reducing cervical cancer, the model places moderate emphasis (0.15 and 0.04) on them compared to the broader health system infrastructure. This implies that cervical cancer prevention strategies should not solely focus on infection rates, as comprehensive healthcare access and long-term investment in vaccination are essential for sustained reductions in cancer rates. Finally, current cost holds a weight of 0.07, since it slightly contributes to the cost-effectiveness of cervical cancer prevention strategies. However, the Extremely Randomized Decision Tree performs worse than the Random Forest, probably because the data points follow an optimized pattern that can be better learned through a regularly designed model (Figure 2).

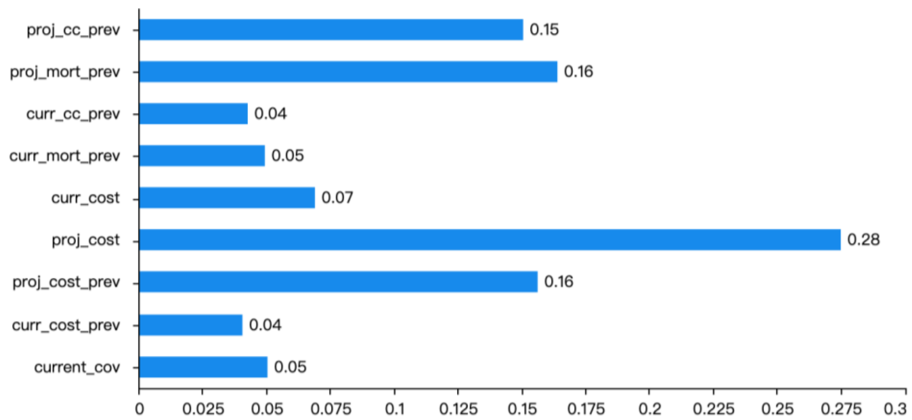

Fig 2. Random Forests (future CC prevention effect).

Furthermore, by applying Gradient-boosted decision trees (GBDT), the author obtained a more optimal model with higher discrimination, which clearly shows that current HPV vaccination coverage is the dominant (weight 0.71) factor influencing the current CC prevention effect, while the weights of other factors take up less than 0.02. Besides, it’s apparent to realize that the current mortality prevention, projective cost, and current are potential predictors, which probably evolve into a decision-making influencing factor in the long term (Figure 2). When applying the same method to a different situation to decide the weights of different factors, the author finds that the projected cost contributes to a weight of 0.55, the projected mortality prevention takes up a weight of 0.28, and the projected cost prevention owns a weight of 0.11. The order of the importance of each predictor is consistent with the results obtained by the previous model. However, the relative gap between different factors has widened even further, which means the second model is more intuitive compared to the previous model. This difference between the effects of the machine learning models depends on the samples they learn with special emphasis based on the difference in their working principles.

By contrast, the application of XGBoost, achieving an optimized effect by implementing a scalable, distributed model based on the GBDT, provides similar results to the ones generated by the GDBT. For the current CC prevention effect, the current HPV vaccination coverage is the primary (weight

0.71) factor. However, for future CC prevention effect (0.36), projected mortality prevention (0.20), projected cost (0.11), and current mortality prevention contribute to a higher proportion compared to their weight in current CC prevention.

4. Conclusion

The author developed a way of categorizing and analyzing the relationship between cervical cancer prevention and vaccination coverage, by classifying data points by the number of projected cervical cancer prevention cases divided by the future vaccination cohort, which includes: (1) 0-25 percentile: very low; (2) 25-50 percentile: low; (3) 50-75 percentile: high; (4) 75-100 percentile: very high.

Therefore, concluding all the processed data listed above, the paper points out the focus of the future cervical cancer prevention, calling for the government to invest in the most influential factors. Current cervical cancer prevention effect, quantified by the number of current cervical cancer prevention cases divided by the current vaccination cohort, is mainly correlated with the HPV vaccination coverage rate. However, future cervical cancer prevention effect, quantified by the number of projected cervical cancer prevention cases divided by the future vaccination cohort, is correlated with the cost paid on HPV vaccination, showcasing the importance of investment in public health care in the long run rather than the short-term HPV vaccination coverage.

Specifically, for future cervical cancer prevention, more factors come into the play, including current cost, current mortality prevention, projected cost, projected cost prevention, and current HPV vaccination coverage. However, the current CC prevention rate contributes 0.66 to the current HPV vaccination coverage.

Comparing the results both from statistical methods and machine learning methods, HPV vaccination coverage remains an important predictor, while the projected cost remains a significant contributor to the cervical cancer prevention. Although the graphs obtained by different machine learning methods reflect a slight difference in weight distribution, the relative priority of different factors remains unchanged. Therefore, although different machine learning models have different degrees of discrimination, the difference is solely caused by the different preferences and tendencies in selecting learning samples and summarizing regular features. In conclusion, although high vaccination coverage can reduce cervical cancer infection rates, the high cost of vaccination still affects vaccine coverage and thus the effectiveness of cervical cancer prevention.

References

[1] Foziya Ahmed Mohammed, et al. Cervical Cancer Stages, Human Papillomavirus Integration, and Malignant Genetic Mutations: Integrative Analysis of Datasets from Four Different Cohorts. Cancers, 2023, 15: 5595-5595.

[2] Canfell Karen, et al. Mortality Impact of Achieving WHO Cervical Cancer Elimination Targets: A Comparative Modelling Analysis in 78 Low-Income and Lower-Middle-Income Countries. The Lancet, 2020, 395: 591-603,

[3] Lehtinen Matti, et al. Eradication of Human Papillomavirus and Elimination of HPV-Related Diseases – Scientific Basis for Global Public Health Policies. Expert Review of Vaccines, 2019, 18: 153-160.

[4] Lehtinen M, et al. Effectiveness of gender-neutral or girls-only vaccination against human papillomavirus. Int J Cancer, 2018, 142(5): 949-958.

[5] Arroyo Mühr, Laila Sara, et al. Concomitant Human Papillomavirus (HPV) Vaccination and Screening for Elimination of HPV and Cervical Cancer. Nature Communications, 2024, 15: 3679.

[6] Hoare Alexander, et al. Sampling and Sensitivity Analyses Tools (SaSAT) for Computational Modelling. Theoretical Biology and Medical Modelling, 2008, 5.

[7] Bhatla N, Albero G, Bruni L. Human Papillomavirus and Related Cancers. Fact Sheet, 2024.

[8] Gibin, Willian oliveira. Cervical Cancer Vaccines. Dataset from Kaggle.com, 2020. www.kaggle.com/datasets/willianoliveiragibin/cervical-cancer-vaccines.

[9] Hanen Karamti, et al. Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach. Cancers, 2023, 15: 4412-4412.

[10] Kalbhor M, Shinde S V, Jude H. Cervical cancer diagnosis based on cytology pap smear image classification using fractional coefficient and machine learning classifiers. Telecommun. Comput. Electron. Control, 2022, 20: 1091-1102.

[11] Youneszade Nina, et al. Deep Learning in Cervical Cancer Diagnosis: Architecture, Opportunities, and Open Research Challenges. IEEE Access, 2023, 6133-6149.