JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A SURVEY OF DESIGN AND DEVELOPMENT OF MACHINE LEARNING MODELS FOR CERVICAL CANCER PREDICTION

GOPU.T [1], HEMA SANDEEPTHI GHANTA [2], NARASIMHA MURTHY CHUNDRU[3], DEEPTHI MALAKALAPALLI[4], THISHITHA SANKURATRI

Asst.Prof [1], UG Student [2,3,4,5]

Department of Computer Science and Technology

Sasi Institute of Technology and Engineering, Andhra Pradesh, India

Abstract : Cervical cancer is a major public health concern, especially in low- and middle-income countries. Lifestyle choices to some extent have an effect on causing cervical cancer. Most cervical cancers are caused by the sexually transmitted infection caused by the Human Papillomavirus (HPV). However, only persistent HPV infections lead to progression to pre-cancer and cancer. The persistence of this infection is influenced by many factors namely, age, sexually transmitted infections, number of sexual partners, age at first sexual intercourse, number of deliveries, tobacco consumption, etc. Risk-based prediction algorithms help to stratify women with a high risk to develop cervical cancer and screen them on a priority basis. In this study, a model has been developed to predict the risk of cervical cancer based on one's lifestyle choices. This is a breakthrough study in the area of women's health that tries to enhance the predictive capability of detection of cervical cancer through machine learning combined with medical diagnosis. ML algorithms like Support Vector Machines, Random Forest, Decision Trees, XGBoost, and Logistic Regression come together to form this ensemble model. The best features of every overall prediction model are used by the ensemble model.

IndexTerms - Machine learning, digital health, cervical cancer, human papillomavirus, risk factors, predictive modelling

I. INTRODUCTION

Though the simulation has shown positive results in testing, the study shows there are some practical problems and the results must be checked cautiously. The results may, in fact, open new avenues for research in the near future with more types of datasets, the latest deep learning techniques, and explainability using AI [1], with a hope that a new era will begin in the treatment of cervical cancer. This study will go further by developing a robust predictive model using state-of-the-art ensemble algorithms with an extensive dataset that includes biomarker, demographic, and clinical data [2]. The aim is to change the way cervical cancer is detected and treated, offering more personalized care, effective faster for those people at risk. Cervical cancer is the seventh most common cancer in the world and ranks as the fourth most common cancer of the female reproductive system. Cancer tumour growth is probably in areas where the cells of the endocervix differentiate into those of the Ex cervix, around the Squamocolumnar Junction too. Terasawa, T. & Hosono, S. (2022). Due to the fact that cervical cancer affects a wide range of demographic groups, it remains a major worldwide health concern. While major strides have been made in both technological developments in screening and in prevention campaigns, the aims of early detection and precise prediction remain thwarted. .[3]. Machine learning, in its present state, is nothing short of phenomenal and constitutes great leaps in deciding upon cervical cancer It will be able to use modern technologies in the form of performance algorithms and deep data mining demographics, from networks, samples, biomarkers, and patient histories. The aim is to open a new frontier in cervical cancer treatment [4]. It proposes a data science approach that pools the predictive powers of Support Vector Machines and Logistic Regression into an ensemble model, Classifier, SVC, and XG Boost as base learners with a Random Forest as a meta learner in a meta-learning framework. In order to do this, a complex strategy is followed that-as it treats the base model's predictions-considers "meta-features," trying to go deeper into the patterns of the data and to push the performance of the models over those relying on techniques with just one model [5]. Key perception develops a new treatment modality approach to cervical cancer, emphasizing personalized medicine using prevention strategies in combination with

precision and efficiency enabled by AI. Projects such as these would be major breakthroughs in the development of novel treatment modalities for cervical cancer and would underscore the importance of timely, focused treatments in the war against cancer today [6]. In this respect, using cutting-edge computing methods and being very careful with the analysis of data, this research will be able to come up with a new bar that catalyzes diagnosis, improving patients' wellbeing. Machine learning has lately been a focus of considerable attention in healthcare, from data analysis and identifying complex patterns to therapeutic innovation. These models can give valuable insights and help in early diagnosis after they have been trained with well-labeled clinical datasets, especially in predicting cervical cancer. The model can thus enhance decision-making by offering speedier Whereas machine learning represents a new paradigm in the conducting of such challenges by equipping computerized systems with the ability to analyze vast volumes of patient data and provide outcome forecasts from the patterns intrinsic in that data [8]. In contrast to classical statistical methods, ML models can handle complex variable interactions with ease to uncover subtle patterns and make predictions at a far greater degree of accuracy when properly trained. The development of machine learning models for cervical cancer prediction offers a promising avenue for improving early detection and reducing the burden of this disease. By combining patient data with advanced computational techniques, we can create predictive tools that assist clinicians in making more informed decisions. This study proposes a predictive model for cervical cancer using logistic regression, decision trees, support vector machines, and random forests. The algorithm selection is guided by data characteristics and the problem at hand, striving for a balance between interpretability of the model and predictive performance [9]. Established metrics such as accuracy, precision, recall, F1-score, and ROC AUC are used to deduce the best model for the prediction of cervical cancer. This slow process also presents a unique opportunity for early diagnosis and treatment. However, traditional screening methods by Pap smears and DNA testing for HPV are also burdened including by a number requirements of disadvantages, for specialized equipment and trained personnel, in addition to inevitable subjectivity of human judgment [10]. These deficiencies create an immediate need for more rapid, easy-to-access, and automated means for predicting cervical diagnostic support that will add efficiency in screening programs [7]. cancer. However, the benefits of a good model in cervical cancer prediction are not limited to the individual patient. A scalable ML-based system can be spread easily at low cost in resource-poor health settings, thereby providing for large-scale screening with minimum requirement of expert intervention [11]. It could reduce the workload of health professionals while ensuring the identification in advance of high-risk patients who will be prioritized for confirmatory diagnosis and treatment.

II. RELATED WORKS

Singh and Sharma 2019 [12] assessed cervical carcinoma detection using a questionnaire study. Performance analysis was done to check if the invention steps in an ANN could be better or worse compared to the other architecture types. ANN has also been employed in determining if a cell is normal, malign, or cancerous. Images of the cervix are taken as a sample to illustrate the methodology for identifying the development of cervical cancer using DT-CWT, which manipulates an input signal within many resolutions, and the oriented local histogram technique thickening edges. D Gupta & Khamparia, 2021[13] propose ML/DL for disease state classification of cancer, breast cancer, COVID-19, physical activity recognition, thermal sensation detection, and dementia assessment for cognitive health. This computation-based system It enables diagnostics with much higher efficiency compared to traditional biochemical approaches, thanks to the big footsteps of medical advancement. Ensemble machine learning along with explainable AI has been used in this work to predict the sequence of events from cervical cancer biopsies. The methods of integer programming and neural networks have been applied correspondingly. Average recall, accuracy, and precision of the ensemble classifier equaled 67%, 94%, and 94% correspondingly. More knowledge and clinical skills were guaranteed. (Jahan et. al., 2021) [14] forecast lethal cancer with the help of ML. Hence, eight algorithms of classification have been developed, by utilizing algorithm building programming. The number one twenty-five important features was used to present the highest performance of the MLP algorithm. It had attained an accuracy and precision among other evaluation matrixes of 97.40% and 97%, respectively, while the recall and f1 score parameters showed a score of 97.4% and 97.2%, correspondingly. T.M. Alam and M.M.A. Khan, 2019 [15] undertook efforts in order to perform tasks which could provide the provisions of additional help materials for the diagnostic procedures that in turn can be employed as a supplementary tool for decision making. Notwithstanding the potential of these models to predict the outcomes related to cervical cancer, these are afflicted with at least one of the following: The drawbacks are the burden of high computational costs, no feature selection technique to reduce dimensionality, lack of resolution approach in addressing imbalance data that may give the model one-side bias, and inadequacy of precision, sensitivity, and specificity of these models. Li, and Chen 2020 [16] employed sequences of photographs taken from colposcopy-view to create a deep learning-based approach that was capable of offering effective diagnosis for both cancer 10 of the cervix as well as CIN. This is a two-tier framework as it utilizes feature matching networks together with key-frame feature encoding networks as its principal two components. In turn, the predictive purposes can be dealt with using machine learning methods such as LR, DT, ANN, SVM, and NB. There is also the possibility of feature optimization techniques, for instance Chicken Swarm optimisation. The prevalence of ML in the present study ranged from the ML application to predict the results of different cervical cancer diagnosis methods. Applications of AI in medicine are aplenty nowadays and cannot be overlooked. It has been proved the extent to which machine learning algorithms pervade all layers of medical data systems and datasets. Fundus image meanings are not confined to glaucoma screening and diabetic retinopathy diagnosis but there are several other purposes as well. The solution was implemented in the form of region-based convolutional neural network technology for cutaneous malignancy recognition. Besides the above, mental health is another important aspect in which AI can be exemplified because it holds great potential and successfully shines. In another investigation, the authors used an ensemble learning technique that considered the risk factors through which the incidence of cervical cancer can be predicted. Indeed, in different research related to the detection of cervical cancer, there is a high instance for the use of deep and machine learning algorithms. However, these algorithms had some constraining which severely limited the model's capability. A partnership of a hybrid learning algorithm is suggested here for a possible answer to challenges imposed by standalone algorithms [17]. A selection of models used was based on a mergence of artificial neural networks algorithms as KNN, CR Tree, Naive Bayes, and SVM. The research obtained the data set both through online acquisition and data pre-processing. The next steps involved uploading the information into the database and then retrieving the basic characteristics. Following this step was the training of algorithms, their implementation on systems relevant for this operation, and their respective validation. The model, during its evaluation period, showed an accuracy of 87.21%. Coming to the ERF, it's this that creates the intersection between the base classifier samples and joining the merger data. This approach suffers

from the major disadvantage of its sensitivity, which is 50% and not sufficient for clinical application. Herly analysis on mean segmentation of k means clustering was done on image pre-processing that included. Authors have extracted from the bulb-shaped nucleus their segmented ground truth data, shape attributes such as convexity and roundness and thickness of the IOU segmentation outcomes. The distinctions have been done by RF classifier based on the parameters of object shape characteristics, while the remaining discriminations have been done by other classifiers. The input data obtained from the UCI cervical cancer dataset were treated by performing classification of its data with three different ML algorithms such as X. Deng, Y. Luo, and C. Wang 2019 [18]. The borderline SMOTE approach identified the imbalances present in the dataset and later got rectified after the implementation of the Borderline SMOTE application. This fact was confirmed through the classifier assessments, as it outperformed SVM by a great margin, such as in the cases of XG Boost and random forest, to separate cancer-malignant from its normal state, benign. There are, given the substantial proportion of missing values in the dataset above, which technique is meant to be used; four possible ways are presented, namely: NOCB-ranked backward carries next, LOCF previous found carried forwarder, filling a median value, and filling a mode value. The authors implemented six ML algorithms to predict the Biopsy target variable: LR, RF, Trees, naïve Bayes, deep learning, and SVM are such linear classifiers. Performance metrics like f1-score and precision-were optimized by the LR and the SVM attendant to NOCB pre processing. It was expressed by Cancer. Net 2019 [19]. However, this did not prevent it from occurring through regular screening or by obtaining the vaccination of HPV. Despite the fact that the vaccine for HPV does appear to epitomize the effectivity of the program in itself, frequent testing remains better apt considering that those above the age bracket of 26 years old are prohibited from being vaccinated. The following two specific tests are thus recommended as screening measures: "Pap cytology also known as Pap test or Pap smears" for some women and "HPV testing.". Abdullah and Fonetta Abdullah. 2019 [20], the list compilation of different gene expressions taking place in precancerous or cancerous cells could manage to hint towards the possibility of discovery. The simplification of model formation process using methods based on machine learning algorithms are being taken forward. Currently, most of the medical diagnosis systems, such as imaging or laboratory tests, use machine intelligence methodologies, and these methodologies have also been used by a range of investigators for the analysis of Pap smear images. However, there are not too many papers that have incorporated machine learning methods into gene expression profiles analysis on the level of cervical cancer.

III. LITERATURE REVIEW

This approach has utilized advanced computational technologies and machine learning methods in an effort to increase the rates of accuracy in detection, hence stimulating timely diagnosis of cervical cancer. Here in, this process takes place through the collection of patient biomaterials relevant to biomarkers, demographic identity, and case histories related to cervical health cancer. Pre processing is that phase of cleaning; this involves the removal of noise, the manipulation of anomalies, and the filling of missing values, a process that increases the consistency of the data accurately.

A. Gathering and Preparing Data

Thus, collection and preparation of data should be very strong for a meaningful analysis, being the first step in trying to build a predictive model to detect cervical cancer. The process is thus a fundamental factor; depending on how well the data and compatibility of the dataset is concerned, the more accurate and reliable the model will be. This project's data was curated from the UCI M.L.R., an eminent repository that holds one of the largest collections of cervical cancer risk factors that result in biopsy investigations. The dataset consists of 858 recorded women volunteers across 36 features. These attributes encapsulate a broad spectrum of factors, possibly indicative of cervical cancer risk, including but not limited to the following: The seattributes encapsulate a broad spectrum of factors, possibly indicative of cervical cancer risk, including but not limited to the following: Demographic Information, Sexual History and Behaviour, Medical History, STD Information, Screening Detection, & Diagnostic Tests, Anomaly Missing Values, Duplicate, Data anonymization, Removal of identifiable information, Conformity to ethical standards. The vigor with which the data collection and preparation was pursued begets a good base on which the rest of the modeling following might take a valuable course. This study will initiate itself as a veritable contributor to the quest for more reliable and ethical data regarding the early detection and prevention of cervical cancer by auditing the quality, relevance, and ethics of the dataset.

B. Data pre-processing method:

The robustness of the predictive model will always depend on the quality of the dataset, as it itself may affect standing. In this respect, the creation of an extensive pipeline of data pre processing, precisely tuned for the cervical cancer dataset available from the UCI Machine Learning Repository, has been a complicated, time-consuming process with great attention to details. This pipeline was designed not only to clean and standardize the data but also improve it so that models performed well Data Type Verification: First of all, pre- processing began excellently by first of all analysing in depth the structure of the dataset: Each column had an accurate classification into being a numerical or categorical. There was extra care taken in demarcation for features that were actually categorical and features that are numerical but their representation is categorical in this dataset. Place Holder Value Treatment: The most frequent issue in the dataset was placeholder values "?" representing missing values: Finding the Placeholders: It scanned through all columns for the presence of "?", indicating where action should be taken. Conversion of Placeholders: Converting numeric values for these columns changed "?" to NaN values, driving consistency throughout the data when it came to missing values. This is a step-by-step pre-processing approach that has laboriously transformed raw data into a structured, rich format from an analytical viewpoint. Handling missing values, standardization of features, and feature engineering have enriched the dataset at this foundation level, so that all further modeling work befits accuracy and insightfulness in the predictive analysis of risk factors for cervical cancer.

C. Consideration of issues

The ethical and legal issues at stake, and their challenges with social dimensions in modeling for the prediction of cervical cancer, are obvious and demand a well-thought-out approach. Implementation of combined approach besides maintaining rigor of research ensures its uptake and adherence by the healthcare system manufacturer. This research has taken into consideration ethical, legal, and social implications of developed technology. It has, therefore, shown its commitment to providing a method that is ethical, legally correct, and socially responsible for the detection of cervical cancer. With such a multi dimensional approach, it not only creates scientific precision and functionality of the predictive model but also alignment with the whole social fabric and healthcare objectives. Therefore, the process of introducing artificial intelligence systems for the earlier detection of cervical cancer through machine learning falls on the platform that males with human rights build trust and provide innovative services.

IV. MODELS TRAINING

After exploration in data analysis, a trial or attempt will be initiated to choose machine learning models that can be used in predicting cervical cancer. In this section, the theoretical background, implementation specifics, and the rationale behind expectations set forth for each model are discussed in view of their contribution toward the goal of increasing cervical cancer detection.

A. Logistic Regression Logistic

Regression determines the likelihood that a queried input is categorized as cervical cancer presence or not accordingly. Also, its output which ranges between zero and 1, is carried out through the logistic function, which is a good choice for binary classification problems. While it does not have as much precision as more detailed models, it can show the level of comparison between different models.

B. Decision Trees Decision

Trees organize instances for their classifying by generating passes through criteria determined on feature values. The tree consists of nodes which are features, branches continuing the child nodes; child nodes representing the possible values of that feature. The leaf of a tree is a class name. This model is opted for its interpretability — since throughout the decision making procedure, the actions follow the same pattern. It is like how humans decide the path to follow. However, the main vulnerability of this model is an overfitting that requires huge amounts of data, as well as methods for their validation.

C. Random Forest

Decision Trees ensemble, multi- random Forest is selected because of stability and accuracy. Summing up many Decision Trees prediction, it reduces the number of factors. It influences the model downhill, improving its accuracy, which falls under overfitting and variance problems.

D. XGBoost

One of the advantages of the XGBoost package is that it is fast in training. Keeping in mind that it is a gradient boosti method, sequentially combines trees such that the first tree 33 corrects errors made by the previous ones as it moves forward. Through the good experience of case studies in winning ML competitions it is possible not only to layout the accuracy for cervical cancer prediction, but also to predict the diseases.

E. Support Vector Machines (SVM)

SVM is partly because of its ability to perform well under conditions of high dimension, even when the dimensions exceed the sample size -- a problem which is very commonly seen on the medical data sets. SVM achieves this by identifying the hyperplane in the feature space which separates classes to the best extent possible, thus providing a powerful classifier. While the kernel trick is helpful for modelling the complex interdependencies which might not be clearly defined in the data, it is also the aspect that makes the machine learning algorithms stand out.

V. MODEL SELECTION CRITERIA

The selection of these models is underpinned by several criteria: interpretability, accuracy, computational efficiency, and their ability to handle imbalanced data. Each model contributes uniquely to the ensemble, offering a diverse perspective on the data. Through comparative analysis and validation. Before model training, the dataset underwent rigorous preprocessing to ensure its suitability for predictive modelling. This involved cleaning, handling missing values through imputation, feature engineering to highlight relevant predictors, normalization of numerical variables, and employing techniques like SMOTE to address class imbalance. These steps were foundational of in preparing the dataset, significantly impacting the models' ability to learn and make accurate predictions.

VI. MODEL ASSESSMENT

VII. This section delves into the systematic evaluation of machine learning models developed to predict cervical cancer, leveraging a dataset encompassing clinical, demographic, and behavioural variables. The assessment focuses on Logistic Regression, Decision Trees, Random Forest, XGBoost, and Support Vector Machines (SVMs), chosen for their suitability in addressing classification problems inherent in medical diagnosis.

• Accuracy

While a fundamental metric, accuracy's usefulness is limited in imbalanced datasets typical of medical diagnosis scenarios. It provides an overall success rate of the model but does not distinguish between the model's performance on the minority and majority classes.

• Precision and Recall

Given the critical nature of medical diagnostics, precision and recall are paramount. Precision illustrates the model's ability to correctly identify positive cases, crucial for minimizing unnecessary anxiety or treatment. Recall, or sensitivity, indicates the model's effectiveness in identifying all actual cases of cervical cancer, essential for ensuring no case goes undetected.

$$Precision = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

$$\text{Recall} = \frac{TruePostive}{TruePostive + FalseNegatives}$$

• AUC-ROC Curve

The AUC-ROC Curve represents the model's ability to discriminate between classes at various threshold levels, providing comprehensive view of performance beyond accuracy. A higher AUC indicates better model performance, crucial for applications like cancer prediction where the cost of misclassification is high.

VIII. RESULTS AND DISCUSSION

This section discusses the results of the feature extraction, oversampling, and the evaluation of the models. The feature extraction process showed significant results. The important features are ranked from most relevant to least relevant as shown in Figure 2. The figure depicts that one feature stands out as tremendously more significant than the others. The 'HPV' feature has the highest feature importance, which corresponds with the real world, i.e. people with HPV are highly at risk of getting cervical cancer. Other than

HPV, other factors namely smoking, age, number of sexual partners, number of pregnancies, and so on are also features that affect the output but they are not as significant as as HPV. This result is clinically accurate and can also be confirmed with the feature correlation heat map. After removing the unimportant features and oversampling using SMOTE, the dataset is used to train a pipeline of various classification models. The models used are Gradient Boosting, XGBoost, Naive Bayes, Ada Boost, Decision Tree, Light GBM, Random Forest, SVM using Radial Basis Function(RBF) kernel, SVM Linear, SVM Polynomial, Multi-Layer Perceptron (MLP), K-Nearest Neighbors(KNN) and Logistic Regression. The results of the models are calculated with and without feature extraction, K-folds.

• Risk Prediction Approach

The risk of developing cervical cancer is primarily calculated using a Gradient Boosting model, which has proven to be one of the most reliable and effective techniques for predicting cervical cancer risk. The Gradient Boosting model, known for its strong performance in classification tasks, has been selected due to its ability to handle non-linear relationships, manage imbalanced data effectively, and deliver high accuracy in predicting outcomes. The model calculates the risk percentage or probability score for each patient based on their input features. These features include a range of lifestyle and clinical factors, such as sexual behavior, smoking habits, age, and, most importantly, whether or not the individual has been diagnosed with Human Papillomavirus (HPV). The presence of HPV is highly correlated with cervical cancer risk, and this factor plays a significant role in the final risk assessment The formula for calculating the probability score is:

$$\text{Probability score} = M+1/N+K$$

Where: M = The number of times a similar instance (with the same feature values) belongs to a particular class (i.e., cervical cancer positive) in the training set. N= The total number of instances in the training set. K= The number of distinct classes in the dataset (e.g., "cancer" and "non-cancer"). This formula ensures that the risk score is adjusted based on the frequency of similar cases in the dataset, thus improving the accuracy of the prediction.

• Detailed Explanation of Risk Factors:

Several factors contribute to the overall risk percentage, including: HPV Status: HPV infection is by far the most significant predictor of cervical cancer. Individuals who test positive for HPV are at a substantially higher risk, and the model reflects this in its calculations.

Smoking Habits: Tobacco use increases the likelihood of cervical cancer, as chemicals in cigarettes can weaken the immune system, making it harder for the body to clear an HPV infection.

Sexual Behavior: The number of sexual partners and the age of first intercourse are also important indicators, as they can increase the risk of contracting HPV and other sexually transmitted infections (STIs).

Reproductive History: The number of pregnancies and deliveries can impact cervical health, and this is factored into the model. Age: Age is another critical factor, as the risk of cervical cancer generally increases with age.

Sample Test Cases for Risk Calculation: To demonstrate how the model predicts the risk percentage for different individuals, let's consider two hypothetical test cases.

Case 1:

Patient Description: An 18-year-old woman.

Relevant Features: Has had five sexual partners.

Smokes 37 packs of cigarettes per year.

Has been diagnosed with sexually transmitted diseases (STDs) three times.

Does not have HPV.

In this scenario, despite the patient's smoking habits and history of STDs, the absence of HPV significantly lowers her risk of developing cervical cancer. The model calculates her risk to be around 19%. This relatively low score reflects the fact that, although she has multiple risk factors, the absence of HPV is a crucial protective factor. However, if this same woman had tested positive for HPV, her risk would dramatically increase to 91.1%. This stark contrast illustrates the overwhelming influence that HPV status has on the overall risk calculation.

Case 2:

Patient Description: A 45-year-old woman.

Relevant Features: Has had only one sexual partner.

Had her first sexual encounter at the age of 30.

Does not have HPV.

In this case, the patient's risk factors are minimal, and the model calculates a risk percentage of only 0.02%. The low number of sexual partners and the late onset of sexual activity, combined with the absence of HPV, result in an extremely low risk of cervical cancer. If, however, this woman had been diagnosed with HPV, her risk of developing cervical cancer would rise dramatically to 80%, despite her otherwise low-risk lifestyle. This case again underscores the critical importance of HPV status in determining cervical cancer risk.

Clinical Relevance of Risk Calculation:

These risk calculations offer more than just a numerical estimate; they provide critical insights into which factors play the most significant role in increasing a woman's risk of cervical cancer. By calculating the risk percentage, healthcare providers can offer more personalized advice and screening recommendations. For example, a woman with a high-risk score, particularly due to HPV, might be advised to undergo more frequent screenings, such as Pap smears or HPV tests, and to consider lifestyle changes that could reduce her overall risk. Similarly, the model's output empowers patients to make informed decisions about their health. Women with higher risk scores may choose to prioritize preventive measures, such as HPV vaccination, smoking cessation, or safer sexual practices. Those with lower risk scores might still benefit from routine screening but may not need to pursue more aggressive interventions.

Impact of the Model on Screening Programs:

This model has the potential to significantly improve the efficiency of cervical cancer screening programs. By identifying women who are at the highest risk of developing cervical cancer, healthcare providers can focus their resources on these individuals, ensuring that they receive priority in screening and follow-up care. This risk-based approach not only helps to prevent the development of cervical cancer but also optimizes healthcare resources by avoiding unnecessary screening for low-risk individuals. Furthermore, the model's ability to assign risk percentages allows for a more nuanced approach to patient care. Rather than a simple binary classification of "cancer" or "no cancer," the model provides a spectrum of risk, giving both patients and providers a better understanding of where they stand and what steps they should take.

Conclusion on Risk Prediction:

The risk prediction model developed in this study demonstrates a high degree of accuracy in predicting cervical cancer risk, particularly by focusing on the influence of HPV status and other lifestyle factors. By calculating individualized risk percentages, the model offers a valuable tool for early detection and prevention efforts, empowering women to take proactive steps in managing their health and enabling healthcare providers to deliver more targeted and effective care.

Conclusion:

Cervical cancer remains one of the leading causes of cancer-related mortality among women globally, particularly in low- and middle-income countries where access to preventive healthcare services is limited. Despite advancements in screening technologies and the widespread availability of vaccines for Human Papillomavirus (HPV), the incidence of cervical cancer continues to be a public health concern. The need for early detection and personalized risk assessment has never been more urgent, especially in regions with limited medical resources. This study contributes to addressing these needs by developing a highly efficient risk prediction model for cervical cancer using advanced machine learning techniques. The proposed model, based on Gradient Boosting and XGBoost, offers an accuracy rate of 98.9%, which outperforms many of the traditional approaches used in earlier studies. The model excels in identifying individuals at a high risk of developing cervical cancer, particularly by focusing on the major contributing factor—HPV status—along with other lifestyle-related risk factors such as smoking, sexual behavior, and reproductive history. The ability to accurately predict the risk of cervical cancer based on these factors has significant implications for both healthcare providers and patients.

REFERENCES:

[1] National Health Portal. (2021) Cervical Cancer. Accessed: Apr. 8, 2023. [Online]. Available: https://www.nhp.gov.in/disease/cancer /cervicalcancer

[2] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, Soerjomataram, and F. Bray, (2020). Global Cancer Observatory: CancerToday.Lyon,France:Internation alAgencyforResearchonCancer. Accessed: Mar. 27, 2023. [Online]. Available: https://gco.iarc.fr/today

[3] WHO Director-General's statement. Who Director-General's Statement on the Call to Eliminate Cervical Cancer as a Public Health Problem. Accessed: Mar. 29, 2023. [Online]. Available: https://www.who.int/initiatives/cervical- cancer-elimination-initiative

[4] N. Al Mudawi and A. Alazeb, ''A model for predicting cervical cancer using machine learning algorithms,'' Sensors, vol. 22, no. 11, p. 4132, May 2022.

[5] N. Kashyap, N. Krishnan, S. Kaur, and S. Ghai, ''Risk factors of cervical cancer: A case-control study,'' Asia–Pacific J. Oncol. Nursing, vol. 6, no. 3, pp. 308–314, Jul. 2019.

[6] S. Dreiseitl and L. Ohno- Machado, ''Logistic regression and artificial neural network classification models: A methodology review,'' J. Biomed. Inform., vol. 35, nos. 5–6, pp. 352–359, Oct. 2002.

[7] J. P. Dsouza, S. Van den Broucke, S. Pattanshetty, and W. Dhoore, ''Exploring the barriers to cervical cancer screening through the lens of implementersandbeneficiariesofthenati onalscreeningprogram:Amulticontextu al study,'' Asian Pacific J. Cancer Prevention (APJCP), vol. 21, no. 8, pp. 2209–2215, Aug. 2020.

[8] H. Alquran, M. Alsalatie, W. A. Mustafa, R. A. Abdi, and A. R. Ismail, ''Cervical Net: A novel cervical cancer classification using feature fusion,'' Bioengineering, vol. 9, no. 10, p. 578, Oct. 2022.

[9] S. Prusty, S. Patnaik, and S. K. Dash, ''SKCV: Stratified K-fold crossvalidation on ML classifiers for predicting cervical cancer,'' Frontiers Nanotechnol., vol. 4, Aug. 2022, Art. no. 972421.

[10] K. Adem, S. Kiliçarslan, and O. Cömert, ''Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification,'' Expert Syst. Appl., vol. 115, pp. 557–564, Jan. 2019.

[11] Bui, B. K. Paul, S. M. Ibrahim, J. M. W. Quinn, and M. A. Moni, ''Machine learning-based statistical analysis for early stage detection of cervical cancer,'' Comput. Biol. Med., vol. 139, Dec. 2021, Art. no. 104985.

[12] A. Juneja, A. Sehgal, A. Mitra, and A. Pandey, ''A survey on risk factors associated with cervical cancer,'' Indian J. Cancer, vol. 40, no. 1, pp. 15–22, 2003.

[13] I. J. Ratul, A. Al-Monsur, B. Tabassum, A. M. Ar-Rafi, M. M. Nishat, and F. Faisal, ''Early risk prediction of cervical cancer: A machine learning approach,'' in Proc. 19th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON). IEEE, May 2022, pp. 1– 4.

[14] M. F. Ijaz, M. Attique, and Y. Son, ''Data-driven cervical cancer prediction model with outlier detection and over-sampling methods,'' Sensors, vol. 20, no. 10, p. 2809, May 2020.

[15] U. K. Lilhore, M. Poongodi, A. Kaur, S. Simaiya, A. D. Algarni, H. Elmannai, V. Vijayakumar, G. B. Tunze, and M. Hamdi, ''Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques,'' Comput. Math. Methods Med., vol. 2022, May 2022, Art. no. 4688327.

[16] W. Yang, X. Gou, T. Xu, X. Yi, and M. Jiang, ''Cervical cancer risk prediction model and analysis of risk factors based on machine learning,'' in Proc. 11th Int. Conf. Bioinf. Biomed. Technol., May 2019, pp. 50–54.

[17] M. B. Rothberg, B. Hu, L. Lipold, S. Schramm, X. W. Jin, A. Sikon, and G. B. Taksler, ''A risk prediction model to allow personalized screening for cervical cancer,'' Cancer Causes Control, vol. 29, no. 3, pp. 297–304, Mar. 2018.

[18] F. Curia, ''Cervical cancer risk prediction with robust ensemble and explainable black boxes method,'' Health Technol., vol.11, no. 4, pp. 875– 885, Jul. 2021.

[19] Y. Li, D. Ge, J. Gu, F. Xu, Q. Zhu, and C. Lu, ''A large cohort study identifying a prediction novel model prognosis for lung adenocarcinoma through machine learning strategies,'' BMC Cancer, vol. 19, no. 1, pp. 1–14, Dec. 2019.

[20] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, ‘‘Machine learning applications in cancer prognosis and prediction,’’ Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, Jan. 2015.