






## Research Article

# Hybrid Model for Detection of Cervical Cancer Using Causal Analysis and Machine Learning Techniques

Umesh Kumar Lilhore <sup>1</sup>, M. Poongodi <sup>2</sup>, Amandeep Kaur <sup>3</sup>, Sarita Simaiya,<sup>3</sup>  
Abeer D. Algarni,<sup>4</sup> Hela Elmannai <sup>4</sup>, V. Vijayakumar,<sup>5</sup> Godwin Brown Tunze <sup>6</sup>,  
and Mounir Hamdi<sup>2</sup>

<sup>1</sup>KIET Group of Institutions, Delhi-NCR, 201206, India

<sup>2</sup>Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

<sup>3</sup>Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

<sup>4</sup>Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

<sup>5</sup>University of New South Wales, Sydney, Australia

<sup>6</sup>Department of Electronics and Telecommunication Engineering, Mbeya University of Science and Technology, Mbeya, Tanzania

Correspondence should be addressed to Umesh Kumar Lilhore; [umesh.lilhore@chitkara.edu.in](mailto:umesh.lilhore@chitkara.edu.in), M. Poongodi; [dr.m.poongodi@gmail.com](mailto:dr.m.poongodi@gmail.com), Amandeep Kaur; [amandeep@chitkara.edu.in](mailto:amandeep@chitkara.edu.in), Hela Elmannai; [hselmannai@pnu.edu.sa](mailto:hselmannai@pnu.edu.sa), and Godwin Brown Tunze; [gtunze@mustnet.ac.tz](mailto:gtunze@mustnet.ac.tz)

Received 29 November 2021; Revised 25 March 2022; Accepted 26 March 2022; Published 4 May 2022

Academic Editor: Po-Hsiang Tsui

Copyright © 2022 Umesh Kumar Lilhore et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The publication of this article was funded by Qatar National Library.

Cervical cancer has become the third most common form of cancer in the in-universe, after the widespread breast cancer. Human papillomavirus risk of infection is linked to the majority of cancer cases. Preventive care, the most expensive way of fighting cancer, can protect about 37% of cancer cases. The Pap smear examination is a standard screening procedure for the initial screening of cervical cancer. However, this manual test procedure generates many false-positive outcomes due to individual errors. Various researchers have extensively investigated machine learning (ML) methods for classifying cervical Pap cells to enhance manual testing. The random forest method is the most popular method for anticipating features from a high-dimensional cancer image dataset. However, the random forest method can get too slow and inefficient for real-time forecasts when too many decision trees are used. This research proposed an efficient feature selection and prediction model for cervical cancer datasets using Boruta analysis and SVM method to deal with this challenge. A Boruta analysis method is used. It is improved from of random forest method and mainly discovers feature subsets from the data source that are significant to assigned classification activity. The proposed model's primary aim is to determine the importance of cervical cancer screening factors for classifying high-risk patients depending on the findings. This research work analyses cervical cancer and various risk factors to help detect cervical cancer. The proposed model Boruta with SVM and various popular ML models are implemented using Python and various performance measuring parameters, i.e., accuracy, precision,  $F1$ -Score, and recall. However, the proposed Boruta analysis with SVM performs outstanding over existing methods.

## 1. Introduction

According to a WHO survey, cervical cancer has probably led to cause cancer affecting women in underdeveloped nations [1]. Despite medical centers, there have been thou-

sands of new cases within the USA in 2016, compared to more than 20K mortality in 2014. This cervical cancer database comprises more than 800 data sample values, 32 characteristics, and four objectives, which have been reported in the year 2016-17. Essential features include aggregate

characteristics, tobacco behaviors, and health records from the past. The several testing and diagnostic procedures that result in an excellent diversity add to the data's complication. As a result, the vital issue involves predicting the person's component behavior and determining the optimum screening technique. As a result, the fundamental problem in predicting the person's component risk assessment is the process of the optimum main channel. Various investigators have examined cervical cancer data collected from different sources [2]. The primary risk factors for cervical cancer transmission are poor menstruation sanitation, adolescent pregnancy, cigarettes, and oral prevention methods. Healthcare datasets have more characteristics and incomplete data than nonmedical datasets. By form of enhancement, it is essential to define the significant and necessary attributes for quantitative model construction. ML techniques are superior in forecasts and performance tuning expeditions, but they have been widely used in cancer and breast cancer research [3]. According to a study [4], long-term HPV infectious disease is the primary cause of cervical cancer.

On the other hand, if diagnosed early and cured correctly, cervical cancer is the most curable type. The technique mentioned above requires more effort to process the information, and obtained low-level features cannot deliver optimal classification efficiency, highlighting the failures of intelligent learning. An ML-based feature extraction approach shares massive advantages over all other cancer detection algorithms in obtaining an improved CAD framework. The ML-based technique accomplishes state-of-the-art findings on complicated computer vision applications [5]. As per existing studies, most cervical precancerous disease classification investigations focus on individual colposcopy visualizations during acetic acid tests, making it challenging to determine cervical cancer. This article focuses on numerous machine learning techniques that can forecast the occurrence of cervical cancer as precisely as feasible, utilizing a fixed number of factors of potential risk determinants for each female. However, the stability of recall and precision is a challenging issue once working to develop a forecasting model with a set of analyses. This research presents a prediction model using machine learning methods to detect cervical cancer analysis. This research proposed an efficient feature selection and prediction model for cervical cancer datasets using Boruta analysis and SVM method to deal with this challenge. This research utilized SVM, random forest, decision tree, and Boruta methods to analyze the cervical cancer dataset. This strongly supports feature classification, regression, clustering, and survival analysis with more modeling methods.

The research work [6] involves the identification of accurate indicators from the UCI dataset that can act as powerful predictors of cervical cancer and a dependent variable that may be a function of these predictions for visualizing and analysis of the cancer trends. Multiple models may be built to find the indicators that can help understand the dynamics of the various variables. The performance of the proposed model and existing ML model is verified using an online cervical cancer dataset using Python and different version mea-

suring parameters, i.e., accuracy, precision,  $F1$  score, and recall. This research is aimed at developing mathematical equations and applying Boruta analysis to depict two types of cervical cancers: (a) low-risk and (b) high-risk cancer. First of all, the cervical cancer dataset has been identified, and the preprocessing has been performed on the dataset, followed by correlation analysis and Boruta analysis. After this, causal analysis has been done that helps identify factors that contribute to cervical cancer. The workflow includes making hypotheses that will be further verified and validated by the results.

The complete research work is organized as follows: Section 1 covers the cancer-related introduction work. Section 2 covers the review of existing research and also suggested a comparative analysis of various methods for cancer research. Similarly, Section 3 covers the materials and techniques, Section 4 covers experiments and results analysis, and finally, Section 5 covers the conclusion and future directions of the research.

## 2. Literature Review

This research presents a machine learning method-based model for earlier cervical cancer prediction in the early stage. This section represents the review of various machine learning models for earlier and more accurately cervical cancer detection. The review work is divided into three subsections based on the risk factor, a mathematical model, and machine learning methods.

*2.1. Based on Risk Factors for Cervical Cancer.* The "National Comprehensive Cancer Network" has issued a warning about the benefits of initial identification of cervical cancer. In contrast, a postponement in treatment is the leading cause of an increasing number of women mortality globally. As a result, numerous scientific and medical investigations have investigated the causes, symptoms, and methodologies of identifying and avoiding cervical cancer. Researchers have also attempted to evaluate the risks that contribute to the pathogenesis and progression of this particular cancer. The selected research works are as follows.

In the research article [7], the cure for cancer has usually taken numerous forms over the years; total elimination may not even be possible; however, the disease's probability of occurrence and forecasting can be reduced. Any disorder can be healed if identified in its beginning phases, and cancer can be successfully treated if spotted in its beginning phases. On the other hand, cervical cancer is hard to forecast in its early stages because there are no symptomatic. The frequent test is done for such forecasting of cancer cells because testing has been the only way it can be forecasted [8]. In [9], to avoid such uncertainties, screening outcomes may be supervised as false positives at points in time, or they may be postponed. Machine learning has been developed in the field of health care services. Numerous methods, techniques, and technology have been used to anticipate cancer cells quicker and with a lower false-positive rate.

The method of mathematical modeling aids in the comprehension of the observable occurrence. The visible event in

TABLE 1: Comparison of a research review on risk factors for cervical cancer.

Article	Risk factors discussed	Imported feature (age group)	Possible cancer types
[15]	Human papilloma-virus (HPV) infection	18-35	Cervical cancer, breast cancer
[16]	Sexual history	Under 18 and above	Carcinoma, cervical cancer
[17]	Smoking	All age groups	Lung, cervical, and breast cancer
[18]	Weakened immune system	30-60	Carcinoma, cervical cancer
[19]	Chlamydia infection	All age groups	Carcinoma, cervical cancer
[20]	Oral contraceptives do with a long period (birth control pills)	18-50	Cervical cancer, lung
[21]	Several full-term pregnancies	18-40	Cervical cancer, lung
[22]	First full-term pregnancy at a young age	25-60	Cervical cancer, lung
[23]	A diet deficient in fruits and veggies	22-56	Cervical cancer, lung
[24]	Smoking and HPV	11-60	Cervical cancer, lung
[25]	Use of pills (pregnancy)	22-45	Cervical cancer, lung
[26]	Early pregnancy, HPV	13-18	Cervical cancer, lung
[27]	HPV and weaker immunity	18-50	Cervical cancer, lung

TABLE 2: Review of cancer type based on no of features and age group.

Article	No of features selected	Imported feature (age group)	Possible cancer types
[33]	13 parameters	18-40	Cancer type 1 and type 2
[34]	10 parameters	20-50	Cervical cancer, lung cancer
[35]	12 parameters	18-55	Cervical cancer, skin cancer
[36]	10 parameters	18-45	Cervical cancer type 3
[37]	15 parameters	20-50	Cervical cancer, breast cancer
[38]	7 parameters	18-30	Cervical cancer, breast cancer
[39]	10 parameters	17-30	Cervical cancer, lung cancer
[40]	18 parameters	14-60	Cervical cancer, type 2 and 3
[41]	12 parameters	15-55	Cervical cancer, type 1

the healthcare area [10] could be wellness symptoms and perhaps a sickness, and this technique results in a workable characterization of complicated things. Inside the medical sciences, the mathematical formulation has also been utilized in various methods to solve, reproduce, research, and explain biological mechanisms [11]. The research [12] proposes probabilistically mathematical systems when the sample sizes are limited and can thoroughly examine the parameters. According to the researchers, any healthcare system may comprehend via comparisons; then, such a procedure must influence the mathematical framework [13]. As illustrated, a model named three separate structures might be used to understand the number of carbohydrates stored in human bodies. Other researchers prefer to use informa-

tive computational methods. These models use a feasible description of factors in analytics testing to describe realistic circumstances [14]. In social and epidemiology investigations, description methods are essential. In most cases, the means, median, average, standard deviation and variance, and other statistics are determined, and a report of the phenomena is written down. Table 1 represents the summary of existing research work based on cancer risk factors.

*2.2. Based on Mathematical Models.* Furthermore, more examination into cervical cancer using mathematical models indicates that significant teams of investigators in the medical sciences concentrate on diagnostics modeling models [28]. The experts in clinical forecasting use a variety of strategies to construct models. Analysis technique and supervised learning model are two examples. Specific healthcare computer models are referred to as “forms of modern.” Basic logical reasoning, hypotheses, concepts, and descriptive analysis have created these frameworks. Many researchers usually refer to such algorithms as medical condition recognition systems [29]. They also utilized ML algorithms to predict serious health issues by the researchers. Enzyme kinetics and pharmacokinetics are two necessary fields of medical research [30]. Machine learning algorithms and automatic analyses are frequently used in several areas of medicine. Physiological reactions and parameters like stress levels, heartbeat, and others must be recorded and modeled for tracking medical conditions within time-series modeling techniques [31]. Modeling, which enables to comprehension of dynamic interaction, uses an approach called transferring characteristics for a detailed look. This type of procedure keeps track of feedback and the processes between this. Many researchers have looked at the principal source of such medical conditions while discovering and establishing the mathematical determinant factors.

Nevertheless, the issue is mainly identifying acceptable factors that can describe the specialized clinical paradigm or phenomenon and determining which independent variables may operate as potential forecasters and which characters can describe the entire computational formula [32]. All

TABLE 3: Comparison of a research review based on machine learning methods.

Article	Technique utilizes	Type of cancer	Important feature discussed	Dataset used	Validation technique
[47]	Artificial neural network	Cancer in breast	Age and mammography results	Diagnostics data and pathological data	Crossvalidation 10-fold
[48]	Support vector machine	Cancer multiple myeloma	STAT1, BRCA1, and CCND1 CCNB1	Online UCI	Crossvalidation 20-fold validation
[49]	Random forest	Cervical cancer	Diet, eating habits, and BME	Clinical data	Crossvalidation 10-fold
[50]	BN methods	Lung cancer	BP, age, and other parameters	Kaggle online dataset	10-fold crossvalidation
[51]	SVM	Cervical cancer, breast cancer	Skin type, breast size, and skin color	Dataset from the hospital (China)	Clinical survey data
[52]	Boruta	Cervical cancer, lung and breast	Age, infection type	Clinical survey data	Crossvalidation
[53]	SVM with random forest	Cervical cancer, cancer in lungs	BME	UCI online dataset	10-fold crossvalidation
[54]	K-NN, SVM	Cervical cancer	Age and mammography results	UCI dataset	Crossvalidation 10-fold

The steps in the Boruta algorithm are as follows:

Step 1: Enhance the data scheme by replicating all factors (so if the original collection has fewer than five features, the data schemes are often prolonged from at least five shadow features).

Step 2: Eliminate the additional features' correlation coefficients with the reaction by shuffling them.

Step 3: On the extensive data system, operate a random forest classifier and collect the Z rankings.

Step 4: Determine the shadow feature with the highest Z score (MZSA), after which allocate a hit to every characteristic that outperformed MZSA.

Step 5: Using the MZSA, initiate a two-test of fairness for every factor of unknown significance.

Step 6: Sign features less importance than MZSA as "insignificant" and eliminates individuals from the data repository forever.

Step 7: Consider the characteristics that have greater significance than MZSA to be "significant."

Step 8: Deactivate all shadow effects.

Step 9: Repeat the above process 9 when all of the characteristics have been allotted significance or the method has achieved the random forest run restriction that was initially established.

ALGORITHM 1: Boruta algorithm.

of the clinical models presented thus far depend on a fundamental grasp of the mathematical model development. Depending on the concerns and obstacles described in the present research, this next section considers the frame of the activity. Table 2 represents the review of cancer types based on several features and age group impact.

**2.3. Based on Machine Learning Models.** In this research, machine learning techniques have been employed to detect cervical cancer accurately via constructing a framework affected by previous research methods in a similar domain. Research [42] proves that by utilizing the oversampling process performance of existing approaches can be improved. This research used the random forest to build a classifier predicated on cervical cancer cases. The analysis indicates that the RF significantly outperformed its same framework after implementing SMOTE, including all characteristics of cervical disease variables in the forms of parameters, i.e., accuracy, specificity, precision, and true positive rate. The research [43] used the online UCI dataset with various strat-

egies for cervical cancer diagnosis: (a) SVM, (b) SVM with PCA, and (3) SVM with RFE. This article concluded that SVM performs well and achieves better precision, diagnostic accuracy, and precision than the multiple different classifiers.

Research [44] utilized three forms of machine learning models to categorize the UCI cervical cancer data. The proposed model used a "border row hierarchical clustering" (BRHC) to deal with dataset inequity. This research has observed that the XG-Boost and random forest methods perform outstandingly in cancer prediction accuracy rates. Since this cancer data contains many incomplete, missing data, it is necessary to deal with missing attributes carefully. Research [45] offers four distinct methods to deal with missing values in the cancer dataset. These techniques are NOCB, LOCF, FVM, and NOCB. To anticipate the biopsy input variables, they utilized six algorithms: LR, RF, SVM, DT, NB, and NN [46], and researchers also concluded that if used with the NOCB preprocessing phase, the SVM, as well as LR, reached the best accuracy,  $F1$  measure, and

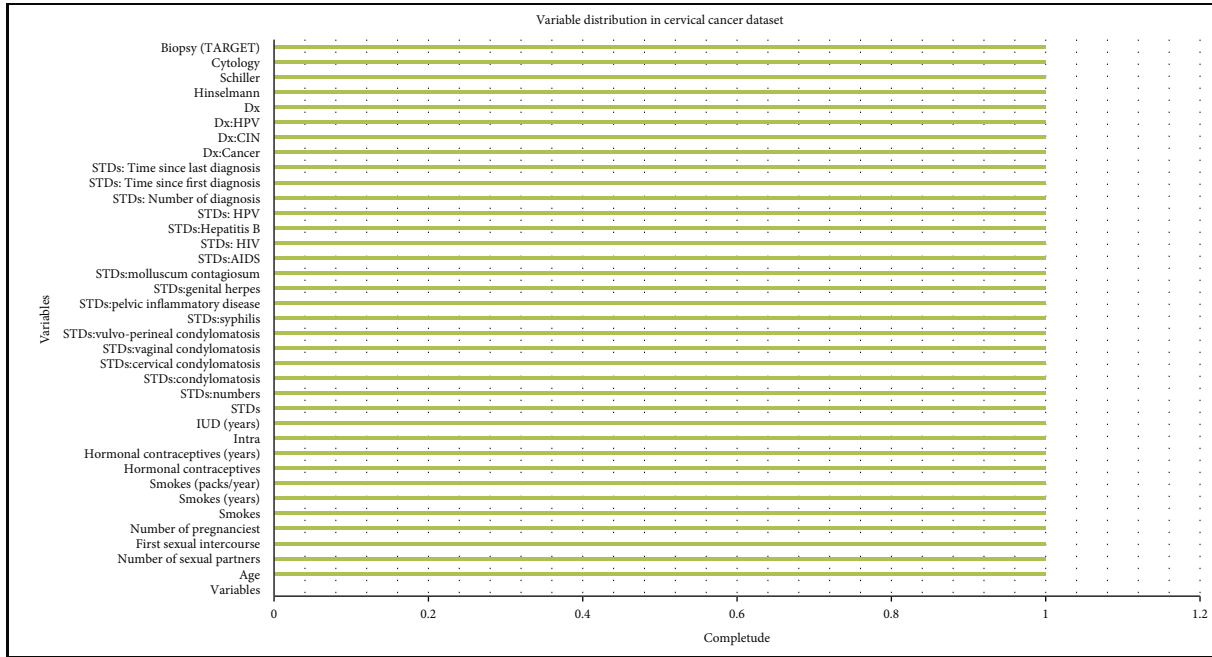


FIGURE 1: Variable distribution in the cervical cancer dataset.

TABLE 4: The hypothesis to find the relationship between the parameters.

S. no.	Dependent parameter	Hypothesis	Possible predictors
1	Sexual partners' frequency	Sexual partners' frequency, Dix: cancer, Dix, STDs: hormonal contraception via hormones (years)	vulvoperinea_lcondy_lomatosis
2	Dix: cancer	Sexual partners' frequency, Dix: cancer, Dix, STDs: hormonal contraception via hormones (years)	vulvoperinea_lcondy_lomatosis
3	STDs: vulvoperinea_lcondy_lomatosis	Sexual partners' frequency, Dix: cancer, Dix, STDs: hormonal contraception via hormones (years),	vulvoperinea_lcondy_lomatosis
4	STDs: condy_lomatosis	Sexual partners' frequency, Dix: cancer, Dix, STDs: contraception via hormones (years),	vulvoperinea_lcondy_lomatosis
5	Contraception via hormones (years)	Sexual partners' frequency, Dix: cancer, Dix, STDs: contraception via hormones (years),	vulvoperinea_lcondy_lomatosis

precision. In this research, machine learning techniques have been employed to detect cervical cancer accurately via constructing a framework affected by previous research methods in a similar domain. The private database was created using 472 survey questions from a China health center, so each cancer patient who took the poll had a correlating gene sequence set of data. This research collects the data from “Mexico’s Maggiore de Caracas health center.” This dataset contains 592 cancer patients’ data with various attributes. This research applied a pooling and discussed the difficulties associated with conventional cervical cancer diagnostics. Table 3 represents the comparison of research methods based on ML methods.

Machine learning approaches have been utilized in this investigation to correctly identify cervical cancer via developing a structure influenced by prior research methodologies used in a similar field. The public available UCI

dataset on cervical cancer does not have per-annotated rows that give a confirmatory signal about the presence or absence of cervical cancer. The dataset aims to understand the subjects that influence a cervical cancer diagnosis.

### 3. Materials and Methods

The section mainly deals with the background research related to the research.

3.1. Predictions of Cancer Risk Factors. Cancer is the second leading cause of death globally, with about 9.6 million deaths in 2019. Cancer is caused when normal cells transform into tumor cells through a multistage process, mainly causing a malignant tumor [55]. However, cancer is more likely to respond to appropriate treatment with an increased chance of survival, less morbidity, and less-expensive therapy if

		Actual	
		0	1
Prediction	0	187	2
	1	25	1

Random forest

		Actual	
		0	1
Prediction	0	186	3
	1	26	0

SVM

		Actual	
		0	1
Prediction	0	188	1
	1	25	1

Decision tree

FIGURE 2: Confusion matrix for random forest, SVM, and decision tree.

identified earlier. Now, it is complex for a computer-aided diagnosis (CAD) system point of view to analyze the complex ecosystem created by screening and diagnosis methods. These complex issues worsen in numerous developing nations due to a lack of computing resources. For all the patients, who are skipping the routine screening, the major problems during diagnosis are identifying the best screening plan and estimating one's risk. The majority of the screening methods correlate with the physician's experience and subjective decision. To determine the riskiest group, one can apply the survey and reduce unnecessary screening. It helps to solve the cancer issues with a plan as per the cancer risk [56]. As per a World Health Organization new survey, cervical cancer has been the "4th greatest common type of cancer." Once especially in comparison to other cancers, this is risky cancer. One such cancer is caused by being infected first alongside the HPV virus [57]. Many scientists discovered that the HPV viral infection is primarily transferred via sexual intercourse. There are many various varieties of HPVs, and cancer has been prompted by category sixteen and pattern 18. These are considered the highest HPVs because they cause cancer cell tissues in the area, so category six and category 11 have been considered significant HPVs because they cause cystitis on the surface [58].

Moreover, it has been found that an efficient and effective detection algorithm was a neural network in the past. The researchers described a TL regularization approach for different linear models, presenting its suitability in various contexts. Positive results have been gathered from this experiment. Other techniques used in cancer detection have been explored, like hierarchical clustering, ANN, and improved genetic algorithms. The authors [59] have performed classification on the cancer dataset, and the results have shown that performance varies between eighty and ninety percent approx. In 2016, the authors [60] had used different data mining techniques and classifiers to predict heart diseases. The researchers have presented the range of performance parameters between forty-five and ninety-nine percent approx. In 2017, the authors [61] did a comparative analysis of different machine learning models utilized for the early detection of heart disease.

**3.2. Machine Learning Methods.** It is a subfield of artificial intelligence (AI) that employs a diverse variety of measurable, statistical inference, and advancement strategies to assist machines in "knowing and understanding" from previous simulation models and comprehending complicated conceptual designs from tremendous, noisy, and complex statistical surveying [62]. Such capacity is helpful for medical

applications that rely on complex proteomic and genotype estimate methodologies. Consequently, intelligence is routinely employed to detect and predict cancerous progression. Machine learning methods have increasingly been designed to estimate and forecast cancer [62].

**3.2.1. Support Vector Machines.** The goal of the model is to find a higher dimensional venue in the  $N$ -dimensional area (where  $N$  represents the total of characteristics that characterize the datasets). Multiple hyperplanes might be used to describe them, but we want to find one with the most significant margin (distance between data points of both classes). Once it is accomplished, future measured values will be able to reinforce and categorized with increased confidence. SVM method creates a hyperplane in a relatively high or infinite space area, which helps in the data categorization process, regression, and other activities, i.e., extracting features and filtering [63].

The hyperplane with the longest distance towards the closest training stage of any category (as such production requires) achieves a better solution because the relatively large the percentage, the reduced the classifier's generalization error message, as described in

$$(X_1, Y_1) \cdots (X_n, Y_n), \quad (1)$$

where  $n$  points and  $X$  and  $Y$  represent the class,  $W$  represents the normal vector, and  $b$  represents the parameter offset of the hyperplane. A hyperplane can be defined as described in

$$W^T(x - b) = 0. \quad (2)$$

**3.2.2. Decision Tree.** DT is a type of nonsupervised learning technique that is commonly utilized for regression and classification problems. The aim is to expand a predictive model of the prediction error using standard decision rules and advanced analytic features [64]. A tree is an example of a fractional estimate. It is represented using the sum of product (SOP) method. Disjunctive normal structure is another name for SOP. So each division out from a massive tree root to just a subtree with the identical class is just a conjunction of attributes, and various branches terminating in that class establish a discontinuity. An entropy  $E$  can be represented as Equation (3).  $E$  represents entropy,  $s$  means samples,  $Py$  represents the probability of yes,  $Pn$  represents no, and  $n$  represents the number of samples.

$$E(s) = \sum_{k=0}^n \binom{n}{k} - Py * \log 2Pn. \quad (3)$$

**3.2.3. Random Forest.** RF is a regression and classification tree-based ensemble learning algorithm. A bootstrap specimen size is used to train each tree, and perhaps optimum solution factors for each separation are chosen from a randomly selected subset of all elements. For regression and classification challenges, the selection processes are distinct. The Gini coefficient was used in the first case, while variance

	Coef	Std err	t	P> t	[0.0251	0.9789]
Intercept	9.936e-17	0.2989	3.43e-17	1.0001	-0.5698	0.5698
BA	0.01008	0.0029	6.9144	0.0019	-0.0088	0.0148
BDA	1.151337	0.1045	14.7899	0.0009	1.3289	1.7147
BMI	0.01279	0.0456	1.0978	0.2789	-0.0108	0.0359
VF	-0.02289	0.0389	-0.6879	0.4977	-0.0889	0.0468
Bpsys	0.04429	0.0345	1.4989	0.1387	-0.0147	0.0998
Bpdia	0.00218	0.0374	0.0998	0.9235	-0.0278	0.0358
SM	-0.00989	0.0301	-0.9357	0.3558	-0.0389	0.0211
Omnibus:		81.7891	Durbin-Watson:			1.7889
Prob (Omnibus):		0.0100	Jarque-Bera(JB):			206.2447
Skew:		0.5248	Prob(JB):			1.65e-45
Kurtosis:		5.2889	Cond. No.			249.19

FIGURE 3: WF having a relationship with other variables.

	Coef	Std err	t	P> t	[0.0251	0.9789]
Intercept	2.266e-19	0.4919	6.98e-18	1.0999	-0.6918	0.7888
BA	0.25088	0.0089	2.6894	0.0479	0.0828	0.0098
BDA	0.41997	0.2458	0.8199	0.4719	-0.2019	0.4088
BMI	0.17911	0.0456	11.8870	0.1059	0.9898	0.3559
WT	-0.01408	0.0399	-0.9979	0.0011	0.1229	0.0868
Bpsys	0.39001	0.0489	13.9989	0.0781	0.4108	0.5198
Bpdia	0.03671	0.1984	2.3481	0.1035	0.0098	0.7124
SM	0.06789	0.0021	5.1157	0.9008	0.4890	0.0891
Omnibus:		19.4911	Durbin-Watson:			1.7977
Prob (Omnibus):		0.0110	Jarque-Bera(JB):			45.9887
Skew:		0.4358	Prob(JB):			2.45e-89
Kurtosis:		3.8989	Cond. No.			241.19

FIGURE 4: Visceral fat (VF) having a relationship with other variables.

	Coef	Std err	t	P> t	[0.0251	0.9789]
Intercept	-2.266e-19	6.8919	-3.58e-18	1.0081	-12.6098	12.6888
VF	1.48088	0.7032	2.9994	0.0889	0.2128	2.8900
BDA	19.00997	2.7258	7.8999	0.0019	14.1889	23.7088
BMI	-0.15809	0.4156	-0.8870	0.6789	-0.7188	0.7759
WT	5.28179	0.8399	6.9979	0.0011	3.8889	6.8568
Bpsys	-2.39429	0.6289	-3.9989	0.0011	-3.9808	-1.9898
Bpdia	0.39718	0.3284	1.9981	0.8035	-0.8908	0.9814
SM	0.00890	0.2801	0.9357	0.9988	-0.8900	0.8991
Omnibus:		19.4911	Durbin-Watson:			1.6977
Prob (Omnibus):		0.0110	Jarque-Bera(JB):			20.2787
Skew:		0.4358	Prob(JB):			3.99e-99
Kurtosis:		3.8989	Cond. No.			99.19

FIGURE 5: BA having a relationship with other variables.

decrease was used in the second case. The RF’s multilateral forecasting has been determined for regression and classification by calculating a majority of votes or an average [65]. The regression method might choose to get a binary result, allowing for probabilistic prediction comparable to regression analysis. The information gain for random forest can be calculated as defined in Equation (4), where  $T$  represents the target variable,  $X$  represents the feature set to be split, and  $\text{Gain}(T, X)$  represents the entropy value after dividing the data feature set  $X$ .

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X). \quad (4)$$

**3.2.4. Boruta Algorithm.** The Boruta method was designed to represent all significant features within a classification model and therefore is designated after a deity of the forest through Slavic mythical. The primary idea behind this method is to use statistical procedures and multiple continuous runs of RFs to evaluate the significance of authentic predictors to arbitrary, as such edge factors. So every run doubles the number of predictors by duplicating them [65]. The shadow explanatory variable model is generated by removing redundant the actual values all over findings, destroying the connection with the results. The different evaluation principles have been accumulated. Compared to a random forest

	Coef	Std err	t	P> t	[0.0251	0.9789]
Intercept	1.286e-19	0.4519	3.52e-18	1.0081	-0.9898	0.9888
VF	0.47708	0.3042	13.8997	0.0889	0.5128	0.9100
BDA	0.24997	0.2458	2.7899	0.0019	-0.1889	0.7088
BMI	0.89809	0.0186	7.3570	0.6789	0.7188	0.2259
WT	0.66178	0.4899	6.9979	0.0011	-0.8889	0.8008
BA	-0.00929	0.0089	-3.9989	0.0011	-0.9808	-1.0098
Bpdia	-0.71018	0.0284	-3.5581	0.8035	-0.8908	0.0114
SM	0.89890	0.0191	0.9357	0.9988	-0.8900	0.0991
Omnibus:		28.7811	Durbin-Watson:			1.3057
Prob (Omnibus):		0.0110	Jarque-Bera(JB):			28.0287
Skew:		-0.4778	Prob(JB):			2.19e-99
Kurtosis:		3.1209	Cond. No.			241.19

FIGURE 6: BPSys having a relationship with other variables.

	Coef	Std err	t	P> t	[0.0251	0.9789]
Intercept	4.146e-17	0.9819	4.87e-19	1.0089	-1.9898	1.9888
VF	0.67708	0.2502	5.0097	0.0889	0.6828	0.9900
BDA	0.89997	0.3988	2.9089	0.0099	0.0889	1.7888
BMI	0.29809	0.0476	4.9070	0.8889	0.0988	0.2889
WT	-0.20178	0.2459	-0.9009	0.8911	-0.4489	0.1208
BA	0.00890	0.0081	0.0857	0.9988	-0.0190	0.8091
Bpdia	0.78929	0.0099	9.9009	0.0011	0.4408	0.0898
BPSys	0.90818	0.0989	8.8081	0.8035	0.7908	0.0914
Omnibus:		48.4011	Durbin-Watson:			1.8957
Prob (Omnibus):		0.0119	Jarque-Bera(JB):			48.8987
Skew:		0.6878	Prob(JB):			8.19e-22
Kurtosis:		3.9909	Cond. No.			240.99

Classification Report					
		Precision	Recall	F1-Score	Support
0		0.91	0.94	0.93	156
1		0.78	0.9	0.88	16
Accuracy				0.95	172
Macro Avg		0.889	0.875	0.757	172
Weighted Avg		0.96	0.95	0.96	172
F1=0.757		AUC=0.555		Recall=0.875	

FIGURE 7: SM has a relationship with other variables.

algorithm mainly learned on the enlarged given dataset, a quantitative test has been conducted for every complex variable; try to reach its significance to the sum of the entire shadow explanatory variable’s maximum values. Algorithm 1 shows the working of Boruta analysis [30, 31].

3.3. *Problem Formulation and Proposed Model.* It is assumed that coefficients can represent the model of cervical detection. The key objective can be understood to be a task(s) of finding an appropriate mathematical model that can be used for cervical cancer causal analysis and mathematically modeling.

There are two tasks involved in finding the changes in the set of variables (independent causal variables  $(X_1, X_2 \dots X_i)$  or single independent variable concerning

the influential variable (dependent variable  $f(y)$ ) that leads to the development of cervical cancer in a subject. Both types of variables share the same vector space model. For a given task  $(T) \longrightarrow \{\{X_1, X_2\}, Y\}$ , the mathematical relationship between these variables is represented by:

$$(T) \longrightarrow f(y) = (\text{Coff}_1 * X_1 + \text{Coff}_2 * X_2 + \text{Coff}_3 * X_3 \dots + \text{Coff}_n * X_n, +C), \tag{5}$$

where  $X_i$  is the set of cervical risk indicators,  $f(y)$  represents the effect that has happened due to  $X_i$ , and  $\text{Coff}_1$  represents the cancer coefficient. We have created a variable, “cervical cancer,” which will be calculated by



$$\text{CervicalCancer} = [\text{Hinselman} + \text{Schiller} + \text{Citology} + \text{Biopsy}]. \quad (6)$$

This research proposed an efficient feature selection and prediction model for cervical cancer datasets using Boruta analysis and SVM method to deal with existing challenges in cervical cancer prediction. A Boruta analysis method is used. It is improved from of random forest method and mainly discovers feature subsets from the data source that are significant to assign classification activity. The proposed model’s primary aim is to determine the importance of cervical cancer screening factors for classifying high-risk patients depending on the findings. Data pre-processing phase plays an essential role in machine learning research because any missing value can affect the entire results. The validity of the data and the essential details that can be extracted significantly influence our model’s potential to gain knowledge; thus, users must pre-process our statistics before supplying them to the proposed model.

### 4. Experiments and Analysis

This section presents the experimental findings and related consequences and discusses the proposed method’s effectiveness over existing methods. This section evaluates numerous practical test parameters for the cervical cancer dataset and compares them with existing ML methods and the proposed methods.

**4.1. Dataset Characteristics.** The dataset consists of 36 attributes representing risk in terms of cervical cancer. Out of these 36 attributes, four attributes are categorical. The values of the categorical attributes are the outcome of the medical tests that have been conducted to verify the clinical finding on cervical cancer. The Hinselman’s test or the colposcopy test is done to check if the lesions are cancerous or not. In Schiller’s test, a part of the body under observation is painted with a solution to investigate the malignant nature of the body part. The cytology test helps ascertain if there is some cancerous fluid in a body part. A complete biopsy is done when most of the standard clinical test options have been exhausted, and only a cut or biopsy can reveal the person’s state of health about cancer, as described in Figure 1. Primary risk variables in constructing a cervical cancer forecasting model include using contraceptive pills, drinking, having many sex partners, and other body parameters.

In summary, the dataset consists of information about lifestyle habits such as smoking, information regarding the sexual behavior of the persons, and, last but not least, about the outcome of the medical tests. It can be observed that the attributes age, number of sexual partners (NSP), HC, and HCY have a correct level of variation, and other attributes’ values do deviate from their mean values. It is because most of these values are Boolean in type. The dataset had a lot of empty values, which requires a missing values’ treatment using the mean and median method.

Classification report

	Precision	Recall	F1-Score	Support
0	0.91	0.94	0.93	156
1	0.78	0.9	0.88	16

Accuracy 0.95 172  
 Macro Avg 0.889 0.875 0.757 172  
 Weighted Avg 0.96 0.95 0.96 172

F1=0.757                      AUC=0.555                      Recall=0.875

FIGURE 8: Experimental results for the random forest method.

Classification report

	Precision	Recall	F1-Score	Support
0	0.901	0.87	0.89	148
1	0.735	0.798	0.875	24

Accuracy 0.91 172  
 Macro Avg 0.845 0.812 0.684 172  
 Weighted Avg 0.912 0.875 0.842 172

F1=0.684                      AUC=0.510                      Recall=0.812

FIGURE 9: Experimental results for the SVM method.

Classification report

	Precision	Recall	F1-Score	Support
0	0.91	0.94	0.93	156
1	0.78	0.9	0.88	16

Accuracy 0.95 172  
 Macro Avg 0.889 0.875 0.757 172  
 Weighted Avg 0.96 0.95 0.96 172

F1=0.757                      AUC=0.555                      Recall=0.875

FIGURE 10: Experimental results for decision methods.

Classification report

	Precision	Recall	F1-Score	Support
0	0.901	0.835	0.725	162
1	0.723	0.758	0.832	10

Accuracy 0.857 172  
 Macro Avg 0.865 0.865 0.718 172  
 Weighted Avg 0.901 0.854 0.825 172

F1=0.718                      AUC=0.534                      Recall=0.865

FIGURE 11: Experimental results for Boruta analysis methods.

**4.2. Results and Hypothesis.** Various machine learning-based models, random forest, SVM, decision tree, and Boruta method have been implemented in Python programming under an anaconda environment.

Table 4 represents the relationship between parameters mainly used for the hypothesis: the dependent parameters and their possible predictors.

**4.2.1. Confusion Matrix.** It is a means of expressing the effectiveness of a classifier’s technique. Once individuals have an inequity number of incidents for each class and when individuals have more than two classes in the data source, a classification performance can be vague (Figure 2).

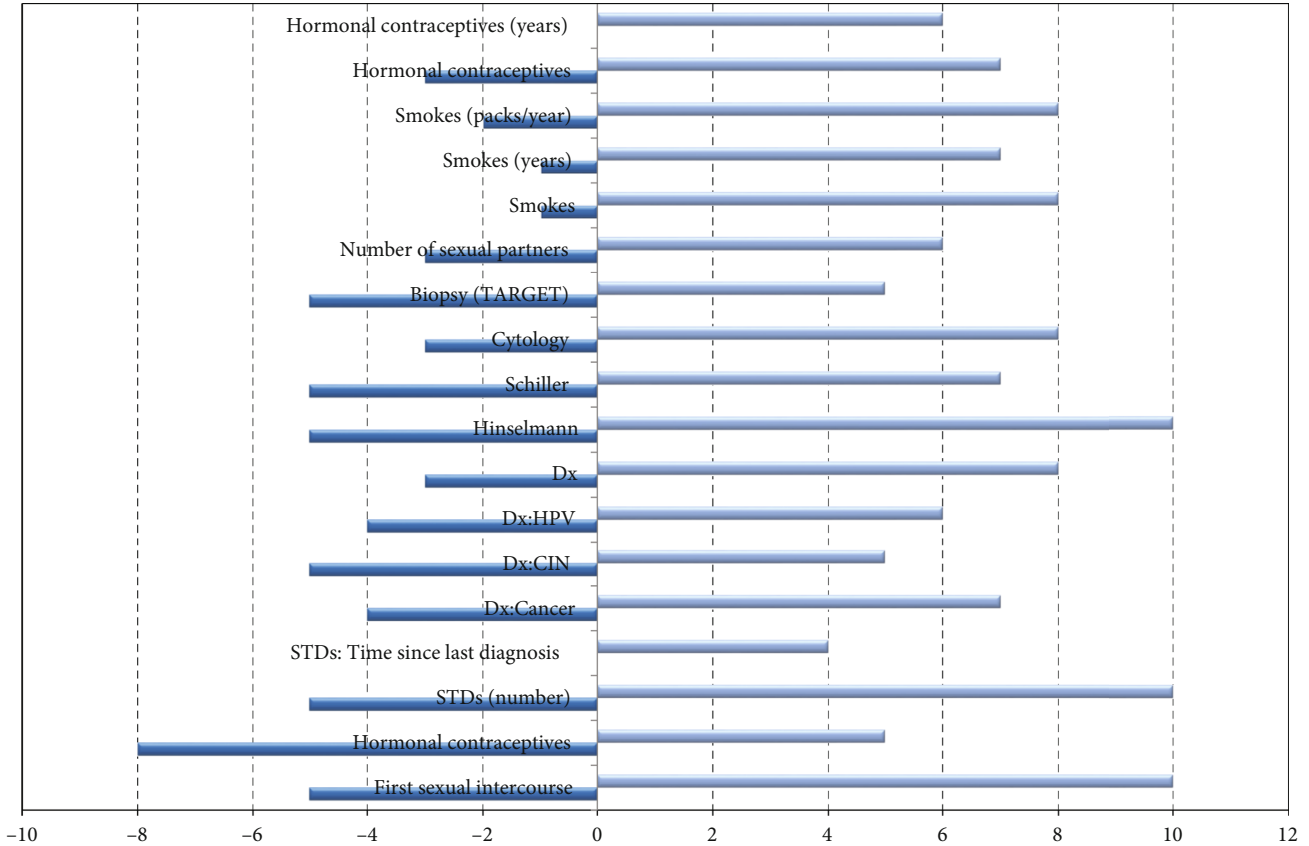


FIGURE 12: Boruta analysis based on selected features.

4.2.2. Experimental Investigation 1

*Hypothesis 1.* Does weight (WT) depend on other parameters or have strong relationships with others.

Interpretation of model 1: Figure 3 shows that the correlation coefficient (coef) varies from  $9.936e-17$  to  $-0.00989$  for different parameters. The coefficient from 0.1 to 0.5 is considered a weak value, and more than 0.5 is considered a substantial value. The  $p$  values show that all parameters are significant as the  $p$  value is ( $p \leq 0.01$ ) for all the variables except in the case of BP-dia, which has a value of 0.99. Moreover, all the parameters have low errors. The intercept has a positive value of 9.93. The data values mainly focus on the mean as SM and VF coefficient values are negative. The total frequency of the findings revealed large tails, indicating that there is no association between dependent class and even the strongest predictor's class of prototype (as described in model 1).

4.2.3. Experimental Investigation 2

*Hypothesis 2.* The value of visceral fat is just a combination of other parameters that directly correlate with others and predictions and verification.

Interpretation of model 2: in Figure 4, the values of coef are less than 0.5, which shows a weak selection. Also, the stan-

dard errors turn out to be close to zero in most cases. At the same time, negative coefficients attribute to it. The  $p$  values are  $\leq 0.01$ , suggesting all the variables' significance and importance. The intercept is positive. The dataset depicts a low level of skewness (shapes are not symmetrical), but massive tails are observed. All these factors fail to acquire the correct coef value.

4.2.4. Experimental Investigation 3

*Hypothesis 3.* This hypothesis mainly considers body age (BA) data and verifies cancer-based on body age. Is the BA a consequence of all the other parameters, or does it have a strong link?

Interpretation of model 3: Figure 5 shows that the  $p$  value for all variables is  $\leq 0.01$ , and it describes that all variables are significant, and this reason stresses including all the significant variables in model 3. The two variables negatively correlate with BA, BMI, and BPsys, whereas all left variables positively correlate with BA. A significant difference was found between the fitted and actual variables' values as they have low standard error except for BDA. The Coef values indicate that the model is not a good fit. The model fails to explain the relationship between BA and other variables.

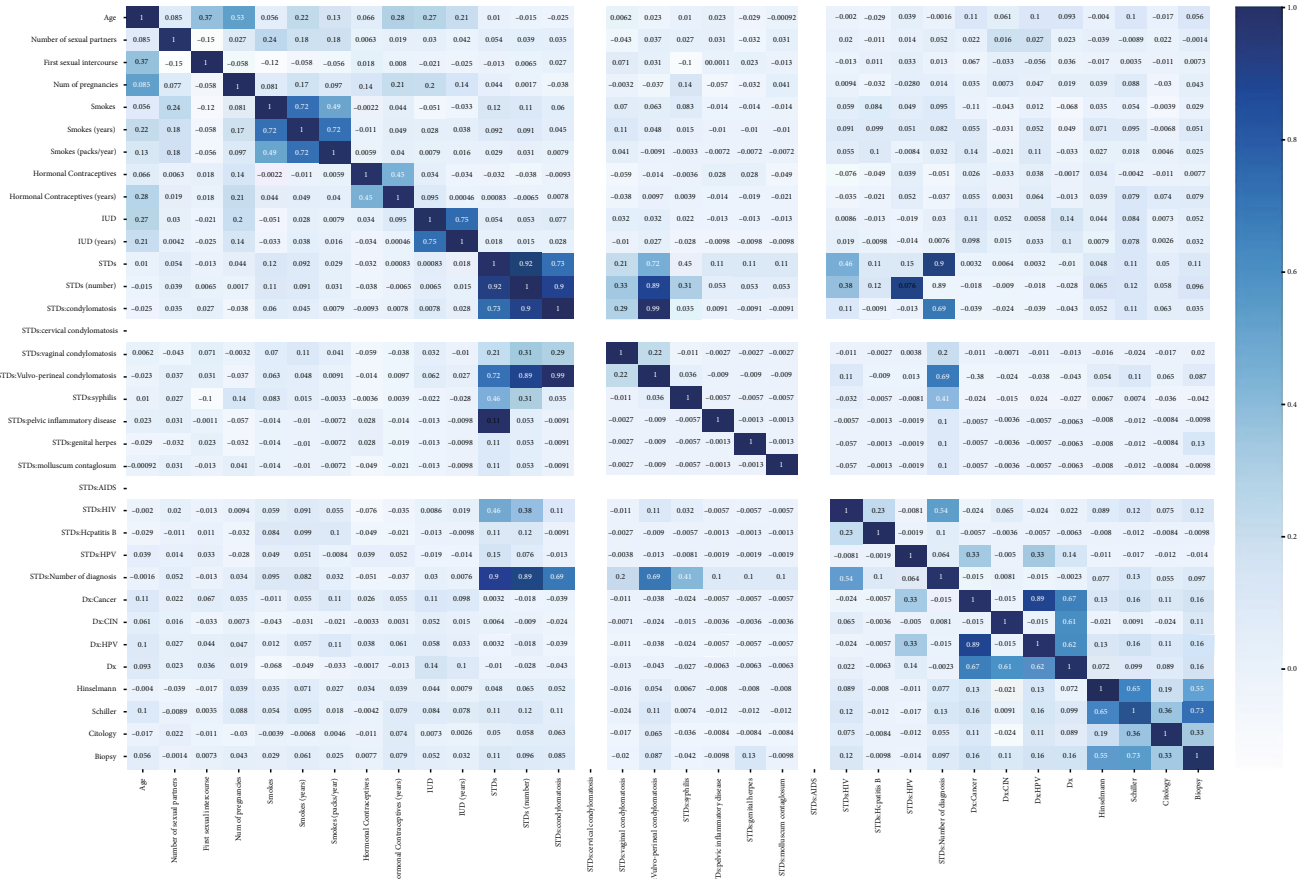


FIGURE 13: Boruta analysis based on all features.

4.2.5. Experimental Investigation 4

*Hypothesis 4.* this hypothesis mainly considers the blood pressure systolic (BpSys) to predict cancer in the body. A BpSys parameter can be used as a function and represent the combination of other parameters with a strong relationship with the other parameters.

Interpretation of model 4: almost all factors have  $p$  values of nil, as shown in Figure 6, which indicates that they are significant and should be incorporated into the equation. Variables BA and BpDia show a minimal negative correlation with the dependent variables (BpSys), whereas a positive correlation is seen with the rest. As a result, we can imply that the regression method had difficulty finding a good fit. It performed pretty finding a good fit as the Coef value (0.754), but the model needs to be rejected with an onset of a better model. Compared with the first model (VF), this model suggests BpSys cannot be a function of all other variables 1.2. The Durbin-Watson test indicates that a high amount of overlap is not desirable.

4.2.6. Experimental Investigation 5

*Hypothesis 5.* This hypothesis mainly considers the skeleton muscle (SM). This hypothesis verifies how the SM param-

eters can be utilized as a cumulative function of other factors with a strong relationship with the numerous parameters.

Interpretation of model 5: the coef is 0.78, which is lower than its VF method designed during the first experiment research. According to the Durbin-Watson results shown in Figure 7, the system has a moderate correlation, indicating that it is just not fit.

4.2.7. Experimental Investigation 6. Hypothesis: in hypothesis 6, we mainly consider the machine learning models. Can the leering machine model with mathematical equations predict cervical cancers accurately? In this experimental investigation, we consider all the hypotheses from 1 to 6 and apply them to various machine learning methods.

Interpretation of model 6: this model utilizes machine learning methods, i.e., random forest, SVM, and decision tree methods. The original data is arbitrarily divided into training and testing pairs to ensure the results obtained are accurate that can be used to create forecasting models. Inside this research work, 70% of the dataset has been used for training, while 30% is used for test results.

The random forest variable's design is directed at the classification method. The overall percentage of vertices inside the RF (the data variable ntree) has been set to 300. Inside the RF method, the total number of trees which will

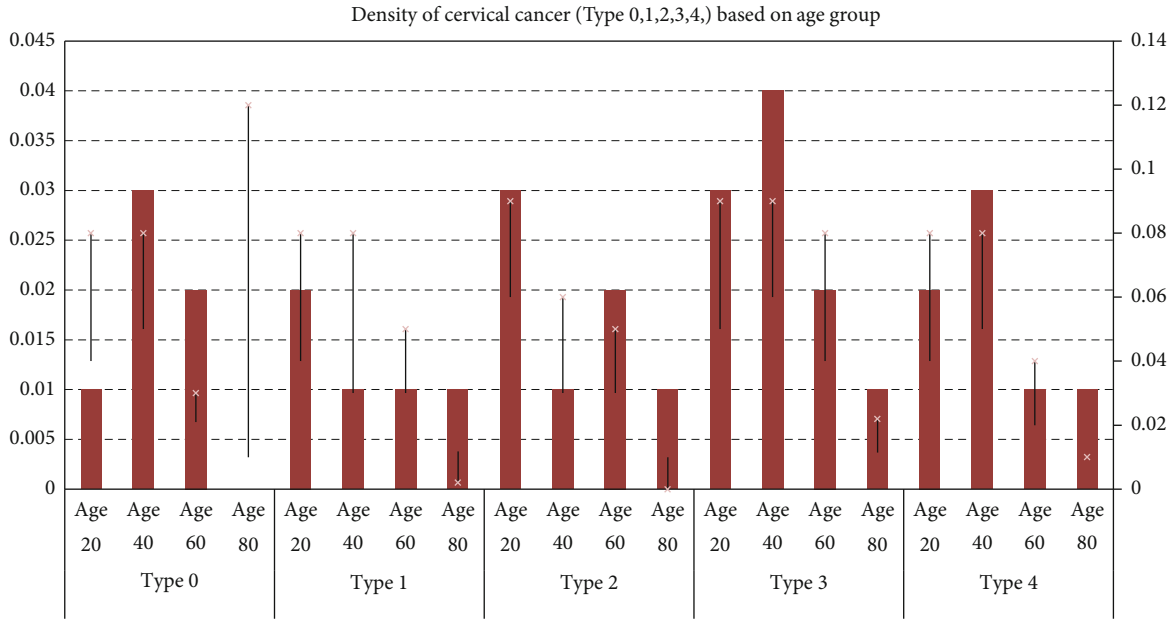


FIGURE 14: Cervical cancer based on age.

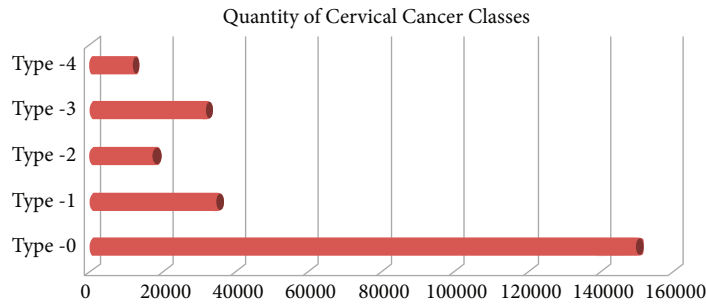


FIGURE 15: Cervical cancer classes.

grow appears to be ntree. We must verify that for almost every source sequence predicted at least very few mins max, the ntree should not be set to a restricted fraction. The study results again for the random forest approach, as shown in Figure 8. In aspects of constructing a predictive model, sixteen samplings have been currently examined for accurate test data. The confusion matrix can have been examined when executing the prediction on the dataset. A confusion matrix can be seen in Figure 6. The confusion matrix will be used to determine how efficient the classifier has been as a prediction. The algorithm anticipated that 7 out of the total eight observations again for normal data sample would be “normal,” although the left standing data sample that was only 1 sample data would be because of cancer. The obtained measurements for SVM approaches are shown in Figure 9.

The precision of the decision tree classifier achieved is greater than 86 percent, which may be appropriate throughout many implementations. In trying to predict cervical cancer, random forest (RF) methods now have one of the highest accuracy appearances. Figure 10 shows the experi-

mental results for the decision tree methods. Figure 11 shows the experimental results for the Boruta analysis methods.

*4.3. Boruta Analysis and Causal Mathematical Modeling Results.* In this section, analyses of all the indicators of cancer are done so that only those variables are used in building an equation model that is useful in detecting cervical cancer. In other words, in this section, the elimination of those variables is done, which does not mathematically correlate to the medical biopsy test. For this purpose, correlation and Boruta importance analysis is done. It is a well-known fact that correlation does not mean a causal relationship between the variables. However, it gives an idea of how strong and weak the relationship is between the variables. Lower correlation values mean the two variables do not have much impact on each other. The Boruta technique evaluated variable importance by swapping predictor qualities and combining them only with initial predictive variables before constructing a random forest upon that fully integrated dataset. After that, we will compare the independent dataset

TABLE 5: Cervical indicators results for Boruta analysis and correlation analysis.

S. no.	Cancer indicator	Boruta analysis	Correlation analysis
	Number of sexual partners	√	√
1	Smoke	√	X
2	Smoke (years)	√	X
3	Smoke (packs)	√	X
4	Hormonal contraceptives (years)	√	X
5	IUD	√	X
6	IUD (years)	√	X
7	STD: number	√	X
8	STD: condylomatosis	√	X
9	STDs: vulvo-perineal condylomatosis	√	√
10	STD: syphilis	√	X
11	STD: time since the first diagnosis	√	X
12	STD: genital herpes	X	√
13	STD: HIV	X	√
14	STD: time since last diagnosis	√	X
15	Dx	√	√
16	Dx_cancer	√	√
17	Dx_HPVP	√	√
18	Dx_CIN	√	X
19	Dx_CIN	X	√
20	Dx_CIN	X	√

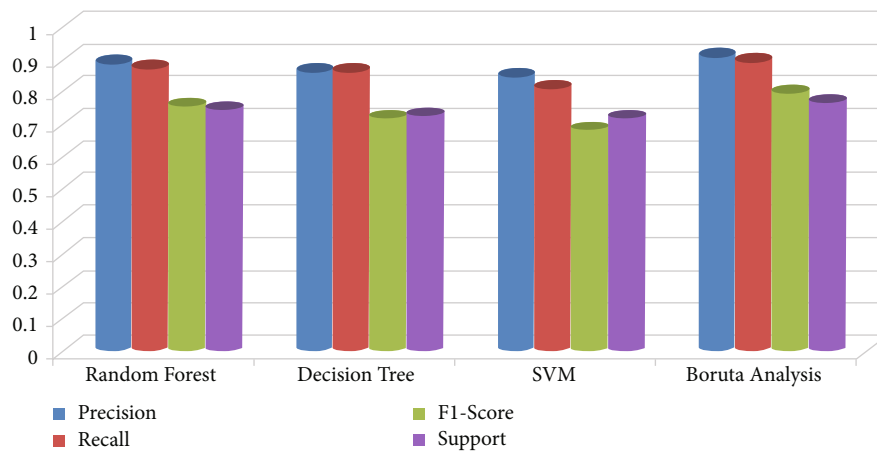


FIGURE 16: Experimental results for Boruta analysis vs. existing methods.

to the randomly selected samples to predict their significance and select something with a higher significance than the randomly selected factors.

According to this graphical analysis (Figure 12), the variables (a) Schiller, (b) Hinselmann, and (c) Cytology have been the most helpful in cervical cancer prediction. Another option for selecting the features is to consider the factors used the most by so many machine learning techniques just to be substantial. Machine learning algorithms first discover the relationship among  $X_s$  and  $Y_s$ , and then, depending on the learning, numerous machine learning methods may utilize multiple parameters to differing extents. As a result, fac-

tors that worked well in a tree-based method like modification or destruction may be undervalued in a linear interpolation model.

As a result, all the factors must be perfectly acceptable for all methodologies. Using an application in a number in ML to select selected features can improve classification performance. Figure 13 shows the Boruta analysis for cervical cancer prediction. It is an algorithm that identifies the importance of the variables for the given categorical variable. This algorithm covers the minimum-optimal feature selection and the all relevant selection strategy. It provides output in terms of three categories, i.e., the most critical variables

and tentatively that are significant during evaluation, and the third category is the rejected features or variables. This algorithm is a wrapper around the random forest algorithm, and its sole purpose is to help select essential variables for further analysis.

Figures 14 and 15 show the details of cervical cancer classes and types (age-based). The cervical cancer classes can be classified into five categories (0, 1, 2, 3, 4, and 5). A correlation histogram showed that two considerations had no other details once all the missing data were filled in (a) sexually transmitted diseases: cervical condylomatosis and (b) sexually transmitted infections (STIs): AIDS. We removed these variables from the dataset and used a comparison heatmap on how each one is connected to the attribute value “tissue sample.” Boruta’s scaling characteristics are the number of characteristics (destroyed) and the collection of instances (correct). So every edge just on the leftmost column refers to a set of particles of the same number of components, so each edge on the top right equates to several characteristics with almost the same number of features. It is worth noting that scalability is sequential concerning the number of features and not so much in terms of the total quantity of particles.

Table 5 represents the results of the cervical indicators for Boruta analysis and correlation analysis. From both kinds of feature analysis, it is clear that “Schiller,” “Hinselmann\_1,” and “Cytology” (medical test) had the highest correlation with biopsy. This means that most medical tests clinically support evidence for cervical cancer. Table 4 gives the output of both these analyses. Logically, some of the attributes out of 36 attributes need to drop. Based on the UCI cervical cancer database, a combination of eighteen characteristics and four diagnostic testing findings are significant for constructing a causality assessment report on cervical cancer. A more profound analysis shows both the methods have found those essential variables: number of sexual partners, Dx: cancer, Dx, STDs: vulvoperineal\_condy\_lomatosis, STD: condy\_lomatosis, hormonal contraceptives (years) are essential.

Hence, it is logical to construct a causal analysis based on these variables. The correlation confirmed that this group of variables is strongly associated. The Boruta algorithm ensures that these variables are significant and vital for further analysis. The analysis confirms the correlation in a few pairs [65]. It is challenging to cover all the dependent and prediction variables due to low correlation values. Then, the section builds a hypothesis around these variables to identify which variable can act as a dependent variable to predict the changes in the dynamics of cervical analysis. Hence, only those variables are used for the subsequent analysis that affects each other and helps predict cervical cancer. Thus, a cervical cancer causal analysis would be formed or nullified by proving a null hypothesis test value. Table 5 gives a set of hypotheses. Multiple performance metrics have been used to enhance the accuracy of clinical overall result forecasting.

Figure 16 shows results for Boruta analysis vs. existing methods. The machine learning methods have been calculated for random forest, SVM, decision tree, and Boruta

analysis on cancer (i.e., cervical cancer) dataset. This research applied ML techniques (random forest, decision tree, SVM, and Boruta analysis) [32] towards cervical cancer prediction and helps in diagnosis to underline the necessity of model development with evidence, considering all the outstanding selected data features such as data cleansing, substituting missing values, and applying a feature extraction approach to increase implications predictions efficiency. This research also utilized ML models to predict the cervical cancer detection risk factors, bearing in mind all the information only within the dataset by substituting variables in the columns by their mean and deleting just the portions with a missing value.

The forecasting results of the models’ coefficient values are near 1, indicating that none of them have reached a high degree of efficiency. In each of the scenarios developed, the diverse range of skills has a substantial effect. Even as  $t$ -test data demonstrated, this correlation between the dependency and independent factors can be completely ruled out. Different scenarios also have expected to be high over 0.76, and the other has a frequency of 0.789 results. The value of cumulative impacts can be calculated as follows:

$$y(\text{VFx}) = [(0.0138 * (\text{BA})) + (0.0811 * (\text{BDA})) - (0.0112 * (\text{WT})) + (0.0128 * (\text{BMI})) - (0.0419 * (\text{BPsys})) - (0.0106 * (\text{BPdia})) + (0.0201 * (\text{SMn}) + 2.05e - 14)]. \quad (7)$$

The experimental values for cervical cancer forecasting using a machine learning algorithm are shown in Figure 16. The obtained Figure 16 measurements are shown for the random forest methodology (precision is 0.889, recall is 0.875,  $F1$  score is 0.757, and support is 0.745). In contrast, the obtained measurements are shown for the decision tree technique (precision is 0.8657, recall is 0.865,  $F1$  score is 0.718, and support is 0.7256). The casual analysis works on the regression process that confirms the statistical relationship between cervical cancer parameters. The results show that “Schiller,” “Hinselmann\_1,” and “Cytology” are the main parameters predicting cervical cancer. When performing superficial root investigation with various parameters, a detailed examination and exploitation of six distinct hypotheses reveal visceral fat represents a healthcare indication and might be a strong predictor of anyone’s health. This parameter indicates that since the rates of other factors include personage, body type, BMI, BP, the metabolism rate, and other essential parameters can represent the correct value of visceral fat. This approach also gives information just on the beginning of medical conditions. The method is verified using multiple measures, including statistical Boruta analysis and correlation, on various machine learning methods.

## 5. Conclusion and Future Work

In this research, a mathematical machine learning-based model has been developed for analyzing various possibilities of cervical cancer. The prediction has been studied by using multiple eight body factors. This research work analyses

cervical cancer and various risk factors contributing to its development. The authors view the statistical technologies, machine learning, and methodologies that can help detect cervical cancer after identifying the paper's research gaps. In addition, this research utilized SVM, random forest, decision tree, and Boruta investigation to create a few classification models. Optimum prospects have been investigated for the development and performance assessment of all modeling techniques. The accuracy and quality of all these methodologies have been analyzed in this article based on the data obtained. Overall, statistical Boruta analysis and random forest methods have performed reasonably well with accuracy, precision, and other parameters for identifying cervical cancer risk and type. The SVM machine learning model produces comparable findings (precision is 0.8456, recall is 0.812,  $F1$  score is 0.684, and support is 0.717). At the same time, the Boruta analysis shows comparable findings (precision is 0.912, recall is 0.891,  $F1$  score is 0.798, and support is 0.768). Compared to other machine learning-based algorithms, the experimental results suggest that Boruta analysis performed best.

Furthermore, this comprehensive evaluation of contouring efficiency may be used to analyze the diagnostic value of fully automated feature extraction in future work. Emerging technologies and methods should be stimulating in research to predict cervical cancer. We can work on socio-demographic factors such as the region of sample data selected and the level of education of that particular region. Educational institutions and schools can contribute to extending the awareness to families of the children they are teaching for their better healthcare.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Acknowledgments

The authors would like to acknowledge Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

## References

- [1] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, and K. G. Kim, "Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images," *Scientific Reports*, vol. 11, no. 1, pp. 16143–16143, 2021.
- [2] C. C. Chang, S. L. Cheng, C. J. Lu, and K. H. Liao, "Prediction of recurrence in patients with cervical cancer using MARS and classification," *International Journal of Machine Learning and Computing*, vol. 3, no. 1, pp. 75–78, 2013.
- [3] M. Poongodi, V. Hamdi, B. S. Vijayakumar, M. Rawal, and Maode, "An effective electronic, waste management solution, based on blockchain smart contract in 5G communities," in *2020 IEEE 3rd 5G World Forum (5GWF)*, pp. 1–6, Bangalore, India, 2020.
- [4] R. Chu, "Risk stratification of early-stage cervical cancer with intermediate-risk factors: model development and validation based on machine learning algorithm," *The Oncologist*, vol. 26, pp. 13956–13956, 2021.
- [5] P. Charoenkwan, W. Shoombuatong, C. Nantasupha, T. Muangmool, P. Suprasert, and K. Charoenkwan, "IPMI: machine learning-aided identification of parametrial invasion in women with early-stage cervical cancer," *Diagnostics*, vol. 11, no. 8, pp. 1454–1454, 2021.
- [6] Z. Zixian, L. Xuning, L. Zhixiang, and H. Hongqiang, "Outburst prediction and influencing factors analysis based on Boruta-Apriori and BO-SVM algorithms," *Journal of Intelligent Fuzzy Systems*, vol. 41, no. 2, pp. 3201–3218, 2021.
- [7] A. Varalakshmi, A. Lakshmi, A. Swetha, and M. Rahema, "A comparative analysis of machine and deep learning models for cervical cancer classification," in *2021 International Conference on System, Computation, Automation, and Networking (ICSCAN)*, pp. 412–425, Puducherry, India, 2021.
- [8] W. Luo, "Predicting cervical cancer outcomes: statistics, images, and machine learning," *Frontiers in Artificial Intelligence*, vol. 4, article 627369, 2021.
- [9] M. Poongodi, V. Vijayakumar, and N. Chilamkurti, "Bitcoin price prediction using ARIMA model," *International Journal of Internet Technology and Secured Transactions*, vol. 10, no. 4, pp. 396–406, 2020.
- [10] A. Jajodia, A. Gupta, H. Prosch et al., "Combination of radiomics and machine learning with diffusion-weighted MR imaging for clinical outcome prognostication in cervical cancer," *Tomography*, vol. 7, no. 3, pp. 344–357, 2021.
- [11] M. M. Patil, "The machine learning algorithm for prediction of risk factors of cervical cancer," *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, no. VI, pp. 4177–4180, 2021.
- [12] P. Kumar, R. Kumar, G. Srivastava et al., "PPSF: a privacy-preserving and secure framework using blockchain-based machine-learning for IoT-driven smart cities," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 2326–2341, 2021.
- [13] M. Poongodi and S. Bose, "Detection and prevention system towards the truth of convergence on decision using Aumann agreement theorem," *Procedia Computer Science*, vol. 50, pp. 244–251, 2015.
- [14] D. Ding, T. Lang, D. Zou et al., "Machine learning-based prediction of survival prognosis in cervical cancer," *BMC Bioinformatics*, vol. 22, no. 1, pp. 331–331, 2021.
- [15] C. Guo, J. Wang, Y. Wang et al., "Novel artificial intelligence machine learning approaches to precisely predict survival and site-specific recurrence in cervical cancer: a multi-institutional study," *Translational Oncology*, vol. 14, no. 5, pp. 101032–101032, 2021.
- [16] M. Poongodi, V. Vijayakumar, B. Rawal et al., "Recommendation model based on trust relations & user credibility," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4057–4064, 2019.
- [17] Q. Yin, "The application of machine learning in cervical cancer prediction," in *2021 6th International Conference on Machine Learning Technologies (ICMLT 2021)*, p. 12, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 2021.

- [18] K. Singh, U. K. Lilhore, and N. Agrawal, "Survey on different tumour detection methods from MR images," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2, no. 5, pp. 589–594, 2017.
- [19] L. Akter, M. M. Ferdib-Al-Islam, M. S. Islam, M. R. Al-Rakhami, and Haque, "Prediction of cervical cancer from behavior risk using machine learning techniques," *SN Computer Science*, vol. 2, no. 3, 2021.
- [20] A. Gupta, A. Anand, and Y. Hasija, "Recall-based machine learning approach for early detection of cervical cancer," in *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–21, Maharashtra, India, 2021.
- [21] C. Prianka and B. Kavida, "Cervical cancer cell prediction using machine learning classification algorithms," *Engineering and Scientific International Journal*, vol. 8, no. 1, pp. 25–29, 2021.
- [22] S. Jahan, M. D. S. Islam, L. Islam et al., "Automated invasive cervical cancer disease detection at early stage through suitable machine learning model," *SN Applied Sciences*, vol. 3, no. 10, 2021.
- [23] A. Arora, Tripathi, and Bhan, "Classification of cervical cancer detection using machine learning algorithms," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pp. 1075–1098, Coimbatore, India, 2021.
- [24] O. Iwendi and A. R. Allen, "Enhanced security technique for wireless sensor network nodes," in *IET Conference on Wireless Sensor Systems (WSS 2012)*, pp. 1–5, London, UK, 2012.
- [25] R. Weegar and K. Sundström, "Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations," *PLoS One*, vol. 15, no. 8, pp. 237911–237911, 2020.
- [26] 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.
- [27] M. Poongodi and S. Bose, "Design of intrusion detection and prevention system (IDPs) using DGSOTFC in collaborative protection networks," in *2013 Fifth International Conference on Advanced Computing (ICoAC)*, pp. 172–178, Chennai, India, 2013.
- [28] A. Hassan, D. Prasad, M. Khurana, U. K. Lilhore, and S. Simaiya, "Integration of internet of things (IoT) in health care industry: an overview of benefits, challenges, and applications," in *Data Science and Innovations for Intelligent Systems*, pp. 165–180, 2021.
- [29] T. R. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," *Malaysian Journal of Computer Science*, pp. 132–148, 2022.
- [30] Z. Ramzan, M. A. Hassan, H. M. S. Asif, and A. Farooq, "A machine learning-based self-risk assessment technique for cervical cancer," *Current Bioinformatics*, vol. 15, 2020.
- [31] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," *Future Generation Computer Systems*, vol. 102, pp. 643–649, 2020.
- [32] A. Kaur and K. S. Mann, "Skeletal bone age assessment using neural network," *International Journal Of Research In Electronics And Computer Engineering*, vol. 5, 2017.
- [33] S. K. Singh and A. Goyal, "Performance analysis of machine learning algorithms for cervical cancer detection," *International Journal of Healthcare Information Systems and Informatics*, vol. 15, no. 2, pp. 1–21, 2020.
- [34] S. Kim, S. Lee, C. H. Choi et al., "Machine learning models to predict survival outcomes according to the surgical approach of primary radical hysterectomy in patients with early cervical cancer," *Cancers*, vol. 13, no. 15, p. 3709, 2021.
- [35] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "COVID-19 pandemic: role of machine learning & deep learning methods in diagnosis," *International Journal of Current Research and Review*, vol. 13, no. 6, pp. 150–155, 2021.
- [36] U. K. Lilhore, S. Simaiya, D. Prasad, and K. Guleria, "A hybrid tumour detection and classification based on machine learning," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 6, pp. 2539–2544, 2020.
- [37] U. K. Lilhore, S. Simaiya, K. Guleria, and D. Prasad, "An efficient load balancing method by using machine learning-based VM distribution and dynamic resource mapping," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 6, pp. 2545–2551, 2020.
- [38] D. T. S. Patel, "A cross sectional study to estimate delay in diagnosis and treatment of tuberculosis (TB) among patients attending urban health centre in an urban slum area," *Public Health Review: International Journal of Public Health Research*, vol. 5, no. 1, pp. 1–7, 2018.
- [39] K. Guleria, A. Sharma, U. K. Lilhore, and D. Prasad, "Breast cancer prediction and classification using supervised learning techniques," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 6, pp. 2519–2522, 2020.
- [40] L. Gupta, A. Edelen, N. Neveu, A. Mishra, C. Mayes, and Y. K. Kim, "Improving surrogate model accuracy for the LCLS-II injector frontend using convolutional neural networks and transfer learning," *Machine Learning: Science and Technology*, vol. 2, no. 4, pp. 1245–1265, 2021.
- [41] D. Y. Fei, O. Almasiri, and A. Rafiq, "Skin cancer detection using support vector machine learning classification based on particle swarm optimization capabilities," *Transactions on Machine Learning and Artificial Intelligence*, vol. 8, no. 4, pp. 1–13, 2020.
- [42] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Machine Learning with Applications*, vol. 6, article 100154, 2021.
- [43] M. S. Ali, M. S. Miah, J. Haque, M. M. Rahman, and M. K. Islam, "An enhanced technique of skin cancer classification using a deep convolutional neural network with transfer learning models," *Machine Learning with Applications*, vol. 5, article 100036, 2021.
- [44] Y. Park, "Classification of cervical cancer using deep learning and machine learning approach," 2021.
- [45] A. M. Abadi, D. U. Department, N. Wustqa, and Nurhayadi, "Diagnosis of brain cancer using radial basis function neural network with singular value decomposition method," *International Journal of Machine Learning and Computing*, vol. 9, no. 4, pp. 527–532, 2019.
- [46] N. Wu, S. Jastrzębski, J. Park, L. Moy, K. Cho, and K. J. Geras, "Improving the ability of deep neural networks to use information from multiple views in breast cancer screening," *The Proceedings of Machine Learning Research*, vol. 121, pp. 827–842, 2020.
- [47] Dataset, "Dataset," <https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29>.



- [48] U. K. Lilhore, A. L. Imoize, C.-C. Lee et al., "Enhanced convolutional neural network model for cassava leaf disease identification and classification," *Mathematics*, vol. 10, no. 4, p. 580, 2022.
- [49] Z. Zhou, G. M. Maquilan, K. Thomas et al., "Quantitative PET imaging and clinical parameters as predictive factors, for patients with cervical- carcinoma: implications of a prediction model generated using multi-objective support vector machine learning," *Technology in Cancer Research & Treatment*, vol. 19, 2020.
- [50] S. Kapil, U. K. Lilhore, and N. Agarwal, "An improved data reduction technique based on KNN & NB with hybrid selection method for effective software bugs triage," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 5, pp. 633–639, 2018.
- [51] M. Kaushik, R. Chandra Joshi, A. S. Kushwah et al., "Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: a machine learning approach," *Computers in Biology and Medicine*, vol. 134, article 104559, 2021.
- [52] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "An efficient credit card fraud detection model based on machine learning methods," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 3414–3424, 2020.
- [53] A. Kaur and K. S. Mann, "A novel framework for cloud-based bone age assessment integration system: review and analysis," *International Journal of Computational Engineering Research*, vol. 7, no. 7, p. 6, 2017.
- [54] S. Sharma, U. Kumar, S. K. Lilhore, N. K. Simaiya, and Trivedi, "An improved random forest algorithm for predicting the COVID-19 pandemic patient health," *Annals of the Romanian Society for Cell Biology*, pp. 67–75, 2021.
- [55] S. Iwendi, J. H. Khan, A. K. Anajemba, F. Bashir, and Noor, "Realizing an efficient IoMT-assisted patient diet recommendation system through machine learning model," *IEEE Access*, vol. 8, pp. 28462–28474, 2020.
- [56] A. Kaur and K. S. Mann, "Segmenting bone parts for bone age assessment using point distribution model and contour modeling," *Journal of Physics: Conference Series*, vol. 933, article 12004, 2018.
- [57] V. Patil and U. K. Lilhore, "A survey on different data mining & machine learning methods for credit card fraud detection," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 5, pp. 320–325, 2018.
- [58] A. Kaur and K. S. Mann, "Hybrid classifier for bone age assessment," *Proceedings of 2nd International Conference on Communication, Computing and Networking ICCCN 2018, NITTTTR*, 2019, pp. 439–444, Chandigarh, India, 2019.
- [59] C. Dhanamjayulu, U. N. Nizhal, P. K. R. Maddikunta, T. R. Gadekallu, and C. Iwendi, "Identification of malnutrition and prediction of BMI from facial images using real-time image processing and machine learning," *IET Image Process*, vol. 16, no. 3, pp. 647–658, 2021.
- [60] Iwendi, "Sanitization: a semantic privacy-preserving framework for unstructured medical datasets," *Journal: Computer Communications*, vol. 161, pp. 160–171, 2020.
- [61] S. Abbas, Z. Jalil, A. R. Javed et al., "BCD-WERT: a novel approach for breast cancer detection using whale optimization-based efficient features and extremely randomized tree algorithm," *PeerJ Computer Science*, vol. 7, 2021.
- [62] S. Simaiya, U. K. Lilhore, D. Prasad, and D. K. Verma, "MRI brain tumour detection & image segmentation by hybrid hierarchical K-means clustering with FCM based machine learning model," *Annals of the Romanian Society for Cell Biology*, pp. 88–94, 2021.
- [63] A. Kaur, M. Khurana, V. Kukreja, P. Jindal, and Geetanjali, "Skeletal growth assessment using segmented middle phalanx with active shape modelling," in *International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 38–41, Greater Noida, India, 2021.
- [64] S. Tiwari, U. Lilhore, and A. Singh, "Artificial neural network and genetic clustering based robust intrusion detection system," *International Journal of Computer Applications*, vol. 179, no. 36, pp. 36–40, 2018.
- [65] A. Kaur and K. S. Mann, "Bone age classification using SVM' international journal of engineering science invention," *International Journal of Engineering Science Invention*, vol. 7, no. 3, pp. 38–45, 2018.