

Available online at <http://www.mecspress.net/ijeme>

Empirical Analysis of Cervical and Breast Cancer Prediction Systems using Classification

¹Prabhjot Kaur*, ²Yashita Pruthi, ³Vidushi Bhatia, ⁴Janmjay Singh

¹Associate Professor, Department of Information technology, MSIT, India.

^{2,3,4}Department of Information technology, MSIT, India.

Received: 30 January 2019; Accepted: 25 April 2019; Published: 08 May 2019

Abstract

Cancer is a life-threatening disease with high mortality rates. In the Indian subcontinent, women have a higher possibility to be diagnosed with cancer than men. The most common cancers identified in Indian women are Breast Cancer and Cervical Cancer. Both these cancers have high survival rates in case of early prediction. This paper reviews the attributes which are used in the existing datasets for prediction of these two cancers. The paper also proposes new attributes to overcome the limitations of existing ones, which will further increase the effectiveness of cancer prediction systems. The efficiency of existing and proposed attributes is compared by processing datasets through data mining algorithms using WEKA tool. The algorithms used for this study are – J48 (Decision Tree), Naïve Bayes, Random Forest, Random Tree, KStar and Bagging Algorithm. The empirical analysis done in the paper reported improvement in the efficiency of cancer prediction over existing prediction systems.

Index Terms: Cancer Prediction Systems, cervical cancer, Breast Cancer.

© 2019 Published by MECS Publisher. Selection and/or peer review under responsibility of the Research Association of Modern Education and Computer Science.

1. Introduction

According to a report by Indian Council of Medical Research – by 2020, 17 lakh more Indians will fall victim to cancer [1]. It is one of the deadliest diseases with limited knowledge of cures. The occurrence of the disease is increasing exponentially worldwide. Chances of survival are more in case of early diagnosis of cancer. Along with endorsing a healthy lifestyle, doctors have recommended [2] regular check-ups and identification of genetic factors and environmental factors. Amongst the Indian population, women are more likely to be diagnosed with cancer than men [3]. The most common cancers identified in Indian women are

* Corresponding author.

E-mail address: ¹thisisprabhjot@gmail.com, ²ypruthi.96@gmail.com, ³vidushibhatia75@gmail.com, ⁴janmjay1111@yahoo.co.in

Breast Cancer and Cervical Cancer [4]. Both the cancers when detected early, promise a high survival rate.

There is sufficient literature written on cancer prediction systems. After reviewing and analyzing the existing studies [5-11], there shortcomings and limitations were observed. This paper focusses on the review of attributes used for the datasets of existing breast and cervical cancer prediction systems and proposes new attributes after identifying the limitation of the existing attributes. Apart from this, statistics on cancer [12, 13] and recommendations of cancer prevention experts are also taken into consideration while proposing new attributes. Comparative analysis of existing and proposed cancer prediction system is done using classification algorithms namely – J48 (Decision Tree), Naïve Bayes, Random Forest, Random Tree, KStar and Bagging Algorithm.

The paper is organized into six sections. Section 2 reviews the attributes of the existing datasets for cancer (breast and cervical cancer) prediction and discuss their limitations. Section 3 discusses the proposed attributes for cancer (breast and cervical cancer) prediction systems. Section 4 briefly explores the classification algorithms which are used to check the efficiency of existing and proposed cancer prediction systems. Section 5 analyses the results derived from the comparative analysis of the datasets formed using existing and proposed systems, followed by the conclusion in section 6.

2. Attributes of Existing Cancer Prediction Systems

For this study, datasets available on trusted online sources [14, 15] are referred. After analyzing the data-sets, based on the quality and quantity of attributes, two datasets with sizeable records and eminent attributes are selected for this study.

2.1. Breast Cancer [16, 17]

The Breast Cancer dataset has 286 records with 85 of them belonging to one class and the rest to other. There are 9 attributes, namely:

- Age: The age of the women.
- Menopause: The age of menopause of the women.
- Tumour Size: The size of tumour recorded in case of cancer.
- Inv-nodes: node size in major portion of breasts
- Node-caps: if the presence of node is in the cap of the breasts
- Deg-malig: stage of breast cancer
- Breast: Affected in left, right or both breasts
- Breast-quad: Specification of location regarding upwards, downwards or towards the center
- Irradiat: Cancer present or not
- Class: Recurrence cases of cancer or non-recurrence cases.

After analysis of this data-set, it is found that some of the parameters mentioned in the data-set are referring to extreme conditions which may not be found in women with early stages of breast cancer. Moreover, some of the parameters cannot be efficiently determined by every woman herself without a medical checkup. Another limitation of this dataset is that it takes into account only those women who are at least diagnosed with one type of cancer and the classification is done with recurrence and non-recurrence cancer cases. Hence, this study observed that there is some gap and the listed attributes of the dataset are insufficient for cancer prediction.

2.2. Cervical Cancer [18]

The cervical cancer dataset has 858 instances distributed on the basis of 36 attributes. One class has 803

instances and the other has 55. The classes are divided on the basis of the biopsy results which include women who were diagnosed positive and negative. The attributes are mentioned below:

- Age: The age of the women
- Number of sexual partners: Number of sexual partners
- First sexual intercourse: Age of first intercourse
- Num of pregnancies: Number of pregnancies.
- Smokes: If the woman is a smoker.
- Smokes (years): Number of years of smoking
- Smokes (packs/year): Frequency of smoking
- Hormonal Contraceptives: Whether there was the use of hormonal contraceptives
- Hormonal Contraceptives (years): Number of years of use.

The presence of following infections and sexually transmitted diseases (STDs) were recorded:

- IUD: intrauterine devices (IUDs)
- IUD (years)
- STDs
- STDs (number)
- STDs:condylomatosis
- STDs:cervical condylomatosis
- STDs:vaginal condylomatosis
- STDs:vulvo-perineal condylomatosis
- STDs:syphilis
- STDs:pelvic inflammatory disease
- STDs:genital herpes
- STDs:molluscum contagiosum
- STDs:AIDS
- STDs:HIV
- STDs:Hepatitis B
- STDs:HPV
- STDs: Number of diagnosis
- Dx:Cancer
- Dx:CIN
- Dx:HPV
- Dx
- Hinselmann
- Schiller
- Cytology
- Class: Diagnosed positive or negative for cancer.

After analysis, it is found that although the dataset is thorough but the women who recorded their data were aware of extensive details about the presence of STDs and HPV (Human papillomavirus). These details cannot be assumed to be in the knowledge of regular women. The major disadvantage of this dataset is that it requires the knowledge of specific attributes which are not commonly known to all women and hence is a limitation in the prediction of cancer disease.

3. Proposed Attributes for Cancer Prediction

3.1. Breast cancer

This paper considered symptoms, statistics and medical recommendations of breast cancer to propose the attributes for breast cancer prediction. The limitations of the existing attributes is corrected to improve the effectiveness of the parameters. The proposed attributes are easy to understand and known to common women without any medical diagnosis. It can be used by any individual to predict cancer. The proposed attributes are:

- (i) **Immediate Family History:** Breast Cancer is highly related to family history. The probability to have breast cancer increases exponentially if an immediate family member has the same. The risk also increases with the number of immediate family members diagnosed. Immediate family members include – mother, sister, daughter, paternal aunt (father’s sister) and grandmother.
- (ii) **Extended Family History:** The effect of the extended family is comparatively less than the effect of immediate family, but it is significant enough to note. It is directly proportional to the risk of breast cancer. Extended family members include – mother’s sisters, cousins, and other male relatives.
- (iii) **Age of Diagnosed Family Members:** If the family members were diagnosed at an early age group like – below 40, then the risk of developing breast cancer increases. Younger the diagnosed family member, higher the risk of breast cancer.
- (iv) **Medical History:** If an individual has suffered from breast cancer earlier then the risk of getting it back is very high. Anyone who has been affected by breast cancer should have regular tests depending on doctor’s recommendations.
- (v) **Age Group:** Statistics have found that older women (age 45 and above) are at a higher risk of breast cancer. Hence, women in this age bracket should get checked more regularly, even if symptoms are not present yet.
- (vi) **Weight:** Women who fall into the category of overweight or obese, according to the BMI index, are highly susceptible to breast cancer and other diseases as well. Along with working towards a healthier lifestyle, regular tests are highly recommended.
- (vii) **Menstrual Cycle:** The higher the difference between menarche and menopause, the higher the risk to develop breast cancer. The women who start menstruation at an early age and have menopause at a much older age (than general) are highly susceptible to breast cancer.
- (viii) **Breast Feeding Period:** Women who have breastfed for a longer period, reduce the risk of developing breast cancer. The threshold period used in this study is 1 to 2 years of breastfeeding.
- (ix) **Estrogen and Progestogen Hormone Therapy:** Hormone therapies can increase the risk of cancer. However, there affect reduces to almost none in 10 years. Therefore, women who have had hormone therapies in the past 10 years are susceptible to breast cancer.
- (x) **Exposure to Radiation:** Women who were exposed to radiation in the chest area (generally at a young age), are at a higher risk of developing breast cancer.
- (xi) **Birth Control Pills:** The effect of birth control pills is analogous to that of hormone therapy. Only women using them in the past 10 years are susceptible to cancer.
- (xii) **Swelling:** Swelling in the chest area is a symptom of breast cancer. Anyone who experiences this, man or woman, are highly recommended to get a medical check-up.
- (xiii) **Irritation:** Irritation in the chest area could be due to many reasons. This is a mild symptom, anyhow, individuals experiencing it for a prolonged time are recommended to get a check-up done.
- (xiv) **Redness:** Redness of the chest area is another mild symptom. This could be due to many reasons but breast cancer is one of them and hence, check-ups are recommended.
- (xv) **Lumps:** Lumps in the breasts are one of the most significant and common symptoms of breast cancer. Anyone who feels a lump in the chest area should get it checked immediately.

- (xvi) **Bleeding:** Bleeding in the chest area of from nodes is a serious symptom. It should be immediately checked by a medical professional.

To review the efficiency of the above attributes in comparison to the existing attributes, a dataset was created. The data is anonymously collected from patients with breast cancer, individuals who have suffered from breast cancer in the past and women who have never suffered from breast cancer with the help of a form (refer Appendix A). A well-formed and diverse train set was generated for prediction of breast cancer with maximum accuracy.

A total of 184 instances were recorded, 42 instances belonged to the class of individuals who were diagnosed with breast cancer and 142 instances of individuals who were not.

3.2. Cervical Cancer

After studying the limitations of the existing parameters for cervical cancer prediction and referring to cervical cancer symptoms [19], statistics [20] and medical recommendations this paper has proposed various attributes for accurate cervical cancer prediction. The proposed attributes are mostly self-diagnosable and can be filled by any individual with nominal knowledge of cancer. The proposed attributes are:

- (i) **Family History:** The probability of acquiring the cancer gene is very high. Hence, anyone with a cancer-diagnosed immediate family member is at high risk and should look out for all the symptoms along with regular check-ups.
- (ii) **Age Group:** Statistics have found that older women (specifically 45-60 years of age) are at a higher risk of cervical cancer [21]. Hence, women in this age bracket should get checked more regularly, even if symptoms are not present yet.
- (iii) **Diethylstilbestrol (DES):** DES was used between the years 1940 and 1971 to prevent miscarriages in pregnant women. It is an artificial estrogen. Due to the harmful side effects, it was banned by the government. However, if an individual's parent has taken DES then the individual can be at risk of cervical cancer.
- (iv) **Human papillomavirus (HPV) Infections:** HPV is one of the most common Sexually Transmitted Diseases (STDs). It is one of the major causes of cervical cancer. Anyone who has suffered from HPV is at risk and should get regular Pap tests.
- (v) **HPV Vaccine:** Individuals who have been vaccinated for HPV have a very less probability of contracting the infection. Hence, it is highly recommended. Moreover, individuals who have been vaccinated have lower probabilities of getting cervical cancer.
- (vi) **Sexually Transmitted Diseases (STDs):** Any other form of STDs can also be a contributing factor in cervical cancer risk. Hence, anyone who has been infected by any STDs is at higher risk of cervical cancer.
- (vii) **Steroids:** Studies have shown that sex steroid hormones have a deep link with the contraception of cervical cancer [22]. Anyone who has had the intake of such steroids is at high risk.
- (viii) **Smoking or any form of Tobacco intake:** Although smoking is directly not related to the development of cervical cancer, for HPV infected women, smoking elongates the HPV infection period and have difficulty in clearing the infection. This increases the risk of the emergence of the cervical precancerous lesion.
- (ix) **Age of becoming sexually active:** Individuals who were sexually active before the age of 18 are at a higher risk of cervical cancer. Anyone who is over 18 and has been sexually active for 12 or more months should get Pap tests done.
- (x) **Multiple sexual partners:** Multiple sexual partners can increase the risk of cervical cancer. This risk is due to the contraception of STDs, especially HPV infection.
- (xi) **Multiple Pregnancies:** Childbirth can also be a reason for cervical cancer. Both vaginal birth and

caesarean birth have chances of leading to cervical cancer. Hormonal changes during pregnancy can make women more susceptible to HPV.

- (xii) **Pap Tests:** Pap tests can indicate precancerous stages of HPV or cervical cancer. Lower stages of cancer indicated in Pap tests are easily cured. Women who are sexually active are recommended to get Pap tests at the frequency suggested by their doctor.
- (xiii) **Abnormal Bleeding:** Abnormal bleeding from the cervix is a very critical symptom of cervical cancer. Pap tests are highly recommended for anyone who experiences abnormal bleeding.
- (xiv) **Pelvic Pain:** Severe pelvic pain can be a mild symptom of cervical cancer. Pelvic pain is a symptom of many other issues and hence, it is not considered a severe one. Nonetheless, individuals who have severe continuous and severe pelvic pain should get themselves checked.
- (xv) **Unusual Discharges:** Unusual discharges from the cervix are a severe symptom of cervical cancer. This symptom indicates infection in the cervical area and could also be an indicator of HPV.
- (xvi) **Frequent or Painful Urination:** Frequent and painful urination is a severe symptom indicating a problem in the reproductive or excretion system of the body. It is also an indicator of HPV infection.
- (xvii) **Unusual Bowel Movement:** Unusual bowel movement is a symptom of many diseases and one of them could be an HPV infection. It can lead to cervical cancer.

To analyze the efficiency of the above attributes, a dataset was created. Anonymous data was collected from patients with cervical cancer and individuals who have never suffered from cervical cancer with the help of a form. The instances were recorded from women of different age groups and varied lifestyle to generate a well-formed and diverse train set for prediction.

A total of 256 instances were recorded, 53 instances belonged to the class of individuals who were diagnosed with cervical cancer and 203 instances of individuals who were not.

4. Classification Algorithms used for the Study

The effectiveness of existing and proposed cancer prediction system is analysed using classification algorithms present in WEKA (Waikato Environment for Knowledge Analysis) tool [23]. It is a popular Java based tool used for classification, regression, data pre-processing, visualization, etc. It consists of machine learning algorithms that can be used for data mining tasks.

Details of the algorithms used in this study are as follows:

4.1. J48 Algorithm [24]

J48 is a decision tree based supervised learning algorithm which uses univariate approach. It is an extension of ID3 and uses divide and conquer to classify data. A decision tree is created by generating different branches according to the choices or paths a problem can take. It is used to conclude the value of a dependent unknown variable on the basis of a one or more independent known values. The limitations of the algorithm is the time complexity due to traversal of longest branch and the space complexity can be high as everything is stored in arrays.

4.2. Naïve Bayes[25]

This algorithm was proposed by Thomas Bayes for supervised learning. Bayesian classifiers allocate the highly likely class to the feature variable. The best results are obtained with functionally dependent features and completely independent features. Its advantage is that it is robust and can tolerate noise in input. The limitation is that it is only capable of checking the presence or absence of a feature.

4.3. Random Forest Algorithm[26]

The algorithm was introduced by Leo Breiman. As the name suggests, it is a combination of tree predictors. These trees are distributed throughout equally. The robust nature and efficiency of a tree decides the error percentage of the algorithm. It can be used for both classification and regression. They have a unique quality of bias reduction which can be used as an advantage by further increasing correlation.

4.4. Random Tree Algorithm [27]

This algorithm considers a tree which is made by various features randomly. These random features are adopted by other probable trees of the same problem. The number of features in these trees at each and every node are equal. All the trees have the same probability of being sampled and hence, on repeating the algorithm enough times on different records, we get accurate results.

4.5. K Star Algorithm[28]

The K star algorithms are lazy algorithms developed by John G. Cleary and Leonard E. Trigg. They store the training records in a look-up table and defer the work as long as possible. Each new record is compared with the older records in order to find relations. The strongest relation i.e. the nearest instance to the current instance is selected and given the same class. It can tolerate noise in input.

4.6. Bagging Algorithm[29]

It combines multiple predictors by sampling multiple train-sets from one train-set. It increases accuracy and identifies bias. With the help of different algorithms for k different train-sets we can have k classifiers. This increases the stability of the regression or classification model. It can also be used in unsupervised cluster analysis. Bootstrap Aggregating is another name for it. A limitation of the algorithm is the high increase in computational complexity.

5. Results and Analysis

To review the performance of the proposed attributes in comparison to the existing attributes, three metrics were recorded – Efficiency, Mean Absolute Error and Root Mean Square Error. The observations were recorded in WEKA with percentage split test option for classifying the dataset into the train set and test set. The percentage breakage of the two sets were 66% and 33% respectively.

5.1. Breast Cancer

The efficiency, Mean absolute error and Root mean square error of the proposed and existing breast cancer prediction systems are listed in Table 1 and are shown graphically in Fig. 1 and 2. It is observed from the figures and Table 1 that efficiency of proposed prediction system is better than the existing system in case of every classification algorithm. Out of all the algorithms, the performance of the proposed system is best with Naïve Bayes.

Table 1. Efficiency, Mean Absolute Error and Root Mean Square Error for existing and proposed Breast Cancer Prediction System

Name of the Algorithm	Efficiency		Mean Absolute Error		Root Mean Square Error	
	Existing System	Proposed System	Existing System	Proposed System	Existing System	Proposed System
J48	68.0412	93.6508	0.3966	0.067	0.4879	0.249
Naïve Bayes	71.134	98.4127	0.3431	0.0351	0.4825	0.1365
Random Forest	70.1031	90.4762	0.37	0.1193	0.4523	0.22
Random Tree	70.1031	93.6508	0.3476	0.0635	0.5275	0.252
KStar	74.2268	92.0635	0.333	0.0861	0.452	0.2605
Bagging	64.9485	88.8889	0.4091	0.14	0.4729	0.2833

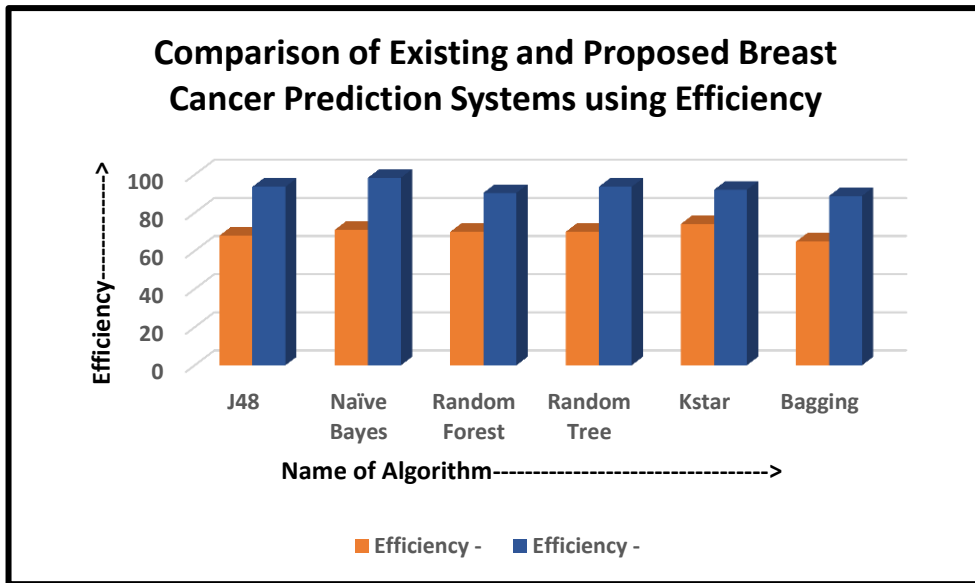


Fig.1. Comparison of Efficiencies of Existing and Proposed Breast Cancer Detection System using Classification algorithms

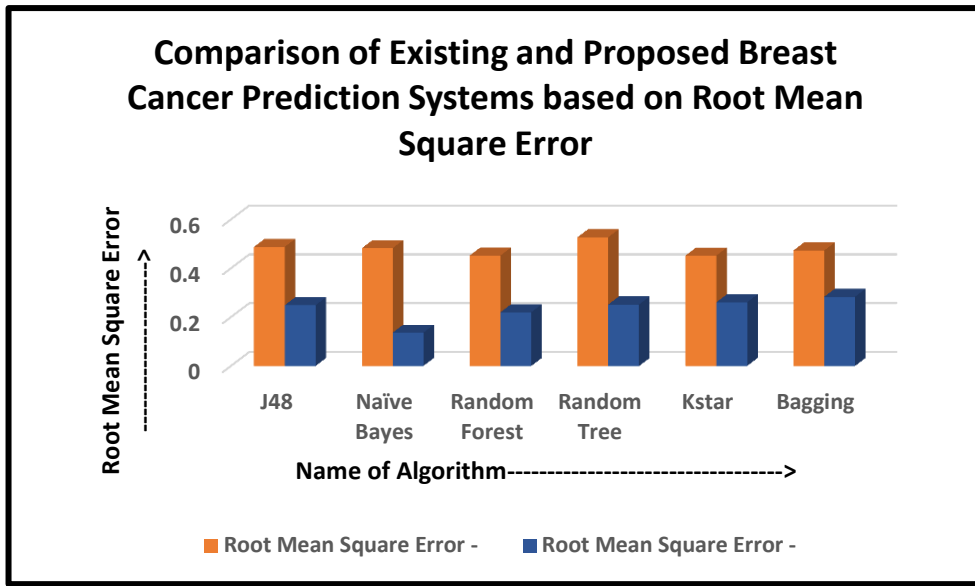


Fig.2. Comparison of Root Mean Square Errors of Existing and Proposed Breast Cancer Detection System using Classification algorithms

5.2. Cervical Cancer

The efficiency, Mean absolute error and Root mean square error of the proposed and existing cervical cancer prediction systems are listed in Table 2 and are shown graphically in Fig. 3 and 4. It is observed from the figures and Table 2 that efficiency of proposed prediction system is better than the existing system in case of every classification algorithm. Out of all the algorithms, the performance of the proposed system is enhanced with the maximum difference of increase in efficiency with Random Tree algorithm.

Table 2. Efficiency, Mean Absolute Error and Root Mean Square Error for existing and proposed Cervical Cancer Prediction System

Name of the Algorithm	Efficiency		Mean Absolute Error		Root Mean Square Error	
	Existing System	Proposed System	Existing System	Proposed System	Existing System	Proposed System
J48	95.189	96.5517	0.0605	0.0345	0.2123	0.1857
Naïve Bayes	85.9107	95.4023	0.1474	0.0467	0.3684	0.2145
Random Forest	95.189	97.7011	0.0654	0.0355	0.1669	0.1518
Random Tree	93.8144	97.7011	0.0619	0.023	0.2487	0.1516
KStar	93.4708	97.7011	0.0698	0.0286	0.2414	0.1522
Bagging	96.3072	96.5517	0.0478	0.0494	0.143	0.1791

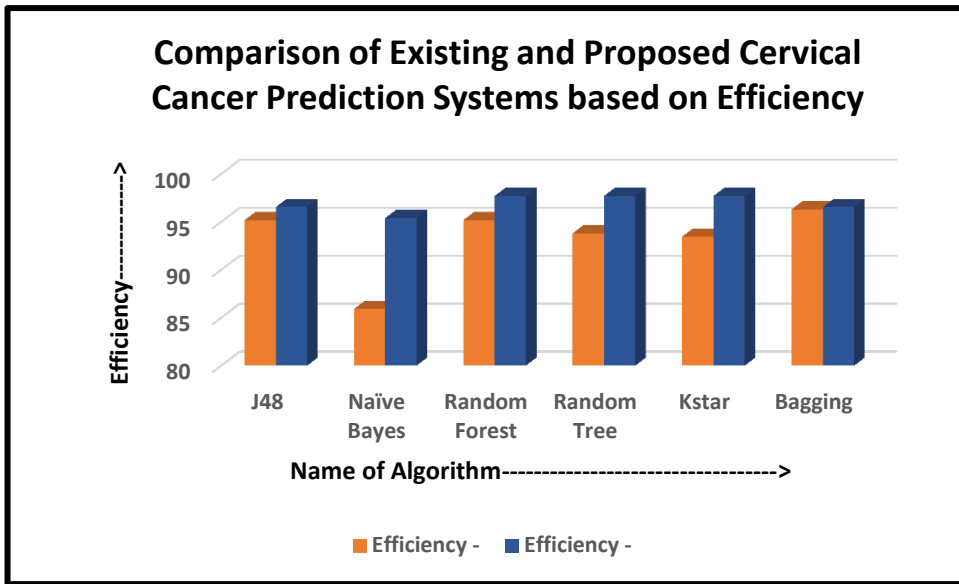


Fig.3. Comparison of Efficiencies of Existing and Proposed Cervical Cancer Detection System using Classification algorithms

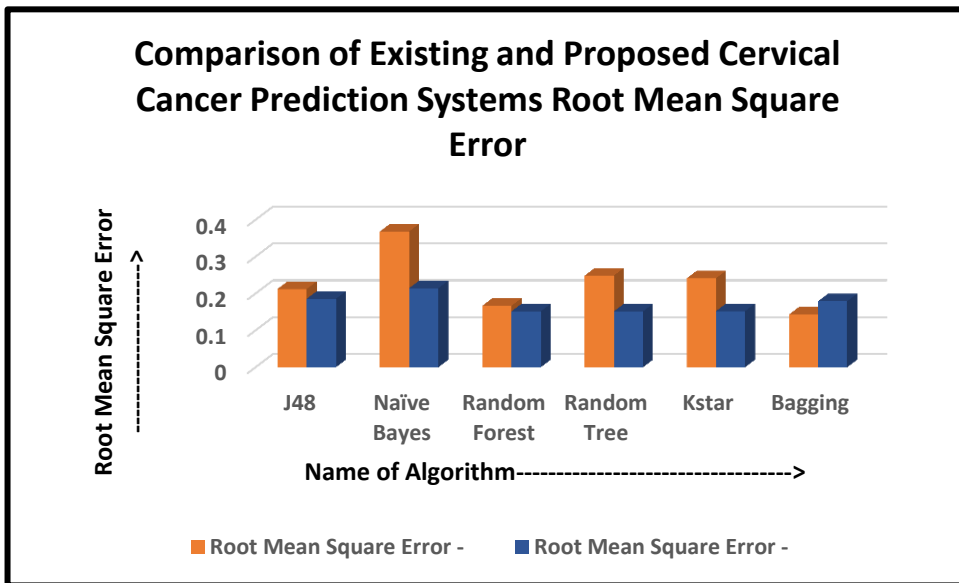


Fig.4. Comparison of Root Mean Square Errors of Existing and Proposed Cervical Cancer Detection System using Classification Algorithms

6. Conclusion

This paper reviewed the existing cancer prediction systems for breast and cervical cancers and proposed new attributes to enhance the performance. The performance of existing and proposed systems was calculated using data mining algorithms namely J48 (Decision Tree), Naïve Bayes, Random Forest, Random Tree, KStar and

Bagging Algorithm using WEKA tool. It is observed from the empirical analysis that proposed system is more efficient than the existing cancer prediction systems.

References

- [1] ICMR (India Council of Medical Research) Report of common cancers in India – <http://www.dailypioneer.com/nation/over-17-lakh-new-cancer-cases-in-india-by-2020-icmr.html>
- [2] Cancer Prevention Measures – <https://www.mayoclinic.org/healthy-lifestyle/adult-health/in-depth/cancer-prevention/art-20044816>
- [3] Cancer statistics based on gender in India, a study by NICPR (National Institute of Cancer Prevention and Research) – <http://cancerindia.org.in/cancer-hits-women-india-men-men-die/>
- [4] Common cancers in India: Research by National Institute of Cancer Prevention and Research. <http://cancerindia.org.in/common-cancers/>
- [5] Dipti N. Punjani, Dr. Kishor H. Atkotiya, “Cervical Cancer Prediction using Data Mining”, International Journal for Research in Applied Science & Engineering Technology, Volume 5 Issue XII, December 2017
- [6] Neelam Singh, Santosh Kumar Singh Bhadauria, “Early Detection of Cancer using Data Mining”, International Journal of Applied Mathematical Sciences, Volume 9, pp. 47-52, 2016
- [7] K. Arutchelvan, Dr. R. Periyasamy, “Cancer Prediction System using Data Mining Technique”, International Research Journal of Engineering and Technology, Volume 2 Issue 8, November 2015
- [8] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra, “Diagnosis of Lung Cancer Prediction System using Data Mining Classification Techniques”, International Journal of Computer Science and Information Technologies, Vol. 4 (1) , pp. 39 – 45, 2013
- [9] A.Priyanga, Dr.S.Prakasam, “The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness”, International Journal of Computer Science and Engineering Communications, Vol.1 Issue.1, December 2013
- [10] A.Priyanga, S.Prakasam, “Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS)”, International Journal of Computer Applications, Volume 83 No 10, December 2013
- [11] P.Ramachandran, N.Girija, T.Bhuvaneshwari, “Early Detection and Prevention of Cancer using Data Mining Techniques”, International Journal of Computer Applications, Volume 97 No.13, July 2014
- [12] General statistics of Cancer – <http://cancerindia.org.in/statistics/>
- [13] Cancer statistics in India from – <http://www.cancerindex.org/India>
- [14] Research source to find multiple datasets – <https://www.kaggle.com/>
- [15] Research source to find multiple datasets – <http://tunedit.org/research>
- [16] Breast Cancer Data Source: <http://tunedit.org/repo/UCI/breast-cancer.arff>
- [17] Breast Cancer attributes description and study – <https://pdfs.semanticscholar.org/4945/4263b6a75a87dbeb94dbe0ba418dba16f459.pdf>
- [18] Cervical Cancer Data Source – <https://www.kaggle.com/loveall/cervical-cancer-risk-classification/data>
- [19] Study on cervical cancer, it's common trends and statistics – <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4404964/>
- [20] Symptoms of cervical cancer – <https://www.cancercenter.com/cervical-cancer/symptoms/>
- [21] Age group of cervical cancer – <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4404964/>
- [22] Sex Steroids and cervical cancer – <https://www.ncbi.nlm.nih.gov/pubmed/22843872>
- [23] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter 11, no. 1 (2009), pp. 10-18.
- [24] Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava, Manish Mathuria, “Decision Tree Analysis on J48 Algorithm for Data Mining”, International Journal of Advanced Research in Computer Science and

Software Engineering, Volume 3, Issue 6, June 2013

- [25] McCallum, Andrew, and Kamal Nigam, "A comparison of event models for naive bayes text classification.", AAAI-98 workshop on learning for text categorization, volume 752, no. 1, pp. 41-48, 1998.
- [26] Leo Breiman, "Consistency for a simple model of random forests", 2004
- [27] Ajay Kumar Mishra, Bikram Kesari Ratha, "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis", International Journal of Electrical and Computer Engineering, Volume 3 Issue 4, 2016
- [28] Dayana C. Tejera Hernández, "An Experimental Study of K* Algorithm", I.J. Information Engineering and Electronic Business, Volume 7, no. 2, March 2015
- [29] G.T. Prasanna Kumari, "A Study Of Bagging And Boosting Approaches To Develop Meta-Classifier", Engineering Science and Technology: An International Journal, Volume 2 Number 5, October 201.

Authors' Profiles



Prabhjot Kaur has completed her B. Tech. in 1999 and M. Tech. in 2003. Presently she is working as an Associate professor in Maharaja Surajmal Institute of Technology, New Delhi, India. Her area of interest includes Soft computing, Image processing, Medical Image segmentation.



Yashita Pruti is pursuing her Bachelor of Information Technology from Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi. Currently, she is doing her major project and this paper is the implementation of her project.



Vidushi Bhatia is pursuing her Bachelor of Information Technology from Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi. Currently, she is doing her major project and this paper is the implementation of her project.



Jannjay Singh is pursuing his Bachelor of Information Technology from Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi. Currently, he is doing his major project and this paper is the implementation of his project.

Appendix A**Forms used to build datasets for proposed attributes**

1. Breast Cancer

<p>1. Please select your age group * <i>Mark only one oval.</i></p> <p><input type="radio"/> Under 30</p> <p><input type="radio"/> Between 30 to 40</p> <p><input type="radio"/> Between 40 to 50</p> <p><input type="radio"/> Above 50</p> <p>2. Weight * Check your BMI here: http://www.nhccworld.in/bmi.html <i>Mark only one oval.</i></p> <p><input type="radio"/> Underweight</p> <p><input type="radio"/> Normal BMI</p> <p><input type="radio"/> Over weight</p> <p><input type="radio"/> Obese</p> <p>3. Started menstruation at the age - Menarche * <i>Mark only one oval.</i></p> <p><input type="radio"/> Before 12 years of age</p> <p><input type="radio"/> After 12 years of age</p> <p>4. Age of Menopause * <i>Mark only one oval.</i></p> <p><input type="radio"/> Haven't had it yet.</p> <p><input type="radio"/> Before 50</p> <p><input type="radio"/> Between 50 and 55</p> <p><input type="radio"/> After 55</p>

5. Breastfeeding *

Have you breastfed? If yes, for how many months?

Mark only one oval.

- Never breastfed
- Less than 6 months
- Between 6 to 12 months
- More than 12 months

6. Undergone estrogen and prostogen hormone therapy *

Have you ever taken part in hormone therapy or any injection of hormones?

Mark only one oval.

- Never Undergone
- Undergone 10 or more years ago
- Undergone within past 10 years

7. Exposure to radiation *

Have you ever had medical (or other) procedures which exposed your chest to radiations?

Mark only one oval.

- Never
- Yes, Once
- Yes, more than once

8. Birth Control Pills *

Have you ever consumed birth control pills?

Mark only one oval.

- Never Consumed
- Consumed 10 or more years ago
- Consumed within past 10 years

9. Medical History *

Have you ever been diagnosed with breast cancer?

Mark only one oval.

- Yes
- No

10. Have you faced any of the following symptoms? (in the chest area)

Mark only one oval per row.


	Yes	No
Swelling	<input type="radio"/>	<input type="radio"/>
Bleeding	<input type="radio"/>	<input type="radio"/>
Lumps	<input type="radio"/>	<input type="radio"/>
Redness	<input type="radio"/>	<input type="radio"/>
Irritation	<input type="radio"/>	<input type="radio"/>

Breast Cancer Tests

16. **Have you been diagnosed with Breast Cancer?**

Mark only one oval.

- Yes, I am currently suffering from breast cancer
- No, I have not tested positive for breast cancer
- I didn't get any tests done.

Powered by
 Google Forms

How to cite this paper: Prabhjot Kaur, Yashita Pruthi, Vidushi Bhatia, Janmjay Singh, "Empirical Analysis of Cervical and Breast Cancer Prediction Systems using Classification", International Journal of Education and Management Engineering(IJEME), Vol.9, No.3, pp.1-15, 2019.DOI: 10.5815/ijeme.2019.03.01