IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.6, June 2020 E-ISSN: 2347-2693

Cervical Cancer prediction based on Hybrid Feature Selection Model and Classification Algorithm

Priyanka Rajpoot[1*], Mahesh Parmar [2]

[1,2]Dept. of CSE & IT, Madhav Institute of Technology and Science, Gwalior (MP) India

**Corresponding Author: priyankarajpoot78047@gmailmail.com*

DOI: https://doi.org/10.26438/ijcse/v8i6.101105 | Available online at: www.ijcseonline.org

Received: 18/Feb/2020, Accepted: 26/Feb/2020, Published: 30/June/2020

Abstract— Cancer has been one of the biggest issues today, The early diagnosis of cancer remains complicated for doctors. When developing novel methods of cancer detection and prevention, It is particularly important to identify genetic and environmental factors. This paper presents the novel approach based on the selection of hybrid features that reduces the dimensionality of features significantly. This paper suggests an efficient Relief and PCA approach that is used on the dataset of cervical cancer. Further, the obtained score is taken as the input for the classification, mechanism. The 3 different classification techniques have been applied. The experiment is conducted on MATLAB. Moreover, The threshold value is experimentally shown to significantly affect the selection of appropriate features. On the basis of many accuracy parameters including accuracy or recall, the experimental result is compared.

Keywords— Hybrid Feature Selection, Chronic Disease Datasets, PCA, classification techniques, Disease Diagnosis, Relief.

I. INTRODUCTION

Cancer(C)'s potential fatal disease is primarily because of environmental factors that transform genes into critical proteins for cell regulation. A resulting abnormal cell activity causes irregular cell masses to spread to vital organ diseases that destroy the surrounding normal tissue, commonly the harbinger of imminent patient death. Most particularly, Globalising unhealthy living, Smoking will increase the incidence of C with the adoption of many modern Western-style foods (low fibre content and high fat). [1].

DM (data method) techniques are developed together to develop a new method for diagnosing certain patient's C. When you start working on a problem of DM. All data must first be combined into the set of instances. The integration of data from different sources is generally difficult. Data to be collected, integrated & cleaned up. Only ML techniques can then be utilized for processing. Doctors and patients may use this developed system to understand the condition and severity of C with ease without screening them for C testing to learn about or treat the disease, large volumes of sensitive information should be recorded and saved.

II. LITERATURE REVIEW

Ritu Chauhan et al [2] Clustering algorithms like HAC & K-Means are based; HAC is used to determine the number of clusters on K-means. When HAC is applied to K-means the quality of the cluster is improved.

Dec hang Chen et al [3] A two-stage cluster method has been developed by EACCD. PAM learns a measure of differences, In the second step, the learned differences are used to obtain the patient clusters with the hierarchical clustering algorithm. The probabilistic system is based on these clusters of patients.

S M Halaani et al [4] Probabilistic algorithms were efficient compared to hierarchical algorithms, all data points are clustered into one cluster, and inappropriate distance measurement choice may be required.

Ada et al [5] An attempt was made to identify lung tumours from cancer images and supportive tools for checking normal & abnormal lungs and to predict abnormal patient survival or years to save a life for patients with cancer.

V. Krishnaiah et al [6] A prototype system for the calculation of lung cancer by DM techniques were developed. Naïve Bayes, followed by the IF-THEN law, seems the most effective approach for predicting patients with lung cancer, Neural Network & Decision Trees. Naïve Bayes observes better results and has been more effective than Decision Trees in the diagnosis of Lung Cancer Disease.

Charles Edeki et al [7] suggested nothing of the algorithms for breast cancer DM & statistical learning exceeded each other to declare an optimal algorithm, and none of the algorithms was badly performed so that survival tasks in breast cancer can be eliminated in the future.

© 2020, IJCSE All Rights Reserved

Zakaria Suliman zubi et al [8] DM methods, like neural networks in X-ray chest films, have been used to detect and classify lung cancer problems to identify features to indicate each of the groups that every case relates.

Labeed K Abdulgafoor et al [9] For intensity-based segmentation, wavelet transformation & K- clustering means algorithms were used.

Sahar A. Mokhtar et al [10] The Decision Tree is analysed for breast mass prediction in three different classification models, support vector machine artificial, neural network.

Rajashree Dash et al [11] The proposed K-means hybrid algorithm combines steps of dimensional reduction through PCA, the new approach to initialization and assigning of data points to appropriate clusters of cluster centre.

III. METHODOLOGY

Several researchers published in the last few Years survivability predictions on cervical cancer and its characteristics. This and survivability prediction using MLT were resolved by the proposed approach for cervical cancer age diseases, Smoking is performed along with other features.

In this research work, three classification algorithms are applied on the Cervical dataset among which Random Forest Classifier outperformed the best. The selection of hybrid features is used as the pre-processing technique in that the hybrid selection method for the diagnosis of cervical cancer shall be associated with relief and PCA methods.

Relief Algorithm

The relief method proposed by Kononenko is one of the most common methods for DM, 1994[32]. It improves on the original relief approach that does not clearly distinguish between redundant properties [40]. The relief method is more reliable & robust or generalizes with multi-class problems. if Euclidean distance is used to find relief technique near hit and near to miss instances, For the same, the Relief uses the distance of Manhattan. Below are the steps to classify weights with the relief algorithm.

1. Start with 0 weights.
2. The example is randomly chosen.
3. Searching for nearest same class neighbours (near-hits).
4. Search for nearby neighbours (near misses) from dissimilar classes.
5. Use the formula of weight estimate based on Step 2 and 3 & 4 values.

Principal Component Analysis (PCA)

PCA is one of the dimensionality techniques used to decrease numbers of datasets as still depicting information about the variability of data. PCA is one of the most well-known DM techniques, extracting major components towards maximum data variability. The first represents the most variance and rest orthogonally with the additional limit that it contains more than possible the maximum variability. The main aspect of PCA is its noise tolerance and its extraction of the strongest data patterns. [39]. Furthermore, In comparison with other techniques for dimensional reduction, it measures maximum data variability.

IV. PROPOSED METHODOLOGY

In the proposed methodology, a combination of relief and PCA is used to develop a method of hybrid feature selection. A most important feature is to choose the optimal threshold value for weights obtained by the relief algorithm. Three classification techniques are applied to the dataset after feature selection to predict the accuracy of the model.

Proposed Algorithm:

Step 1: Load Cervical Cancer Dataset

Step 2: The imported data is pre-processed.

Step 3: Apply the following selection of hybrid feature to pre-processed Data :

- Apply weight relief & rank of every data set feature.
  - Using weight retrieves from relief to set thresholds.
  - Create the list of features above the threshold with weights.
- Use PCA on the above list.

Step 4: The performance evaluation depends on the time & percentage of selected features.

Step 5: Apply classifiers on the obtained score

Step 6: Run the algorithm

Step 7: Obtain the predicted results

Based on each feature's weight, all data set attributes are assigned a rank. The next step is to set the threshold for the number of features to be selected for each data set. The relief algorithm with different thresholds will be tested. Finally, the system presented with Relief Method removes the feature under the threshold value specified or lists features that satisfy a threshold value specified. The key component analysis method applies after applying the Relief algorithm that assigns a score to every selected feature. This technique is used with different thresholds, mean median & standard deviation.

© 2020, IJCSE All Rights Reserved

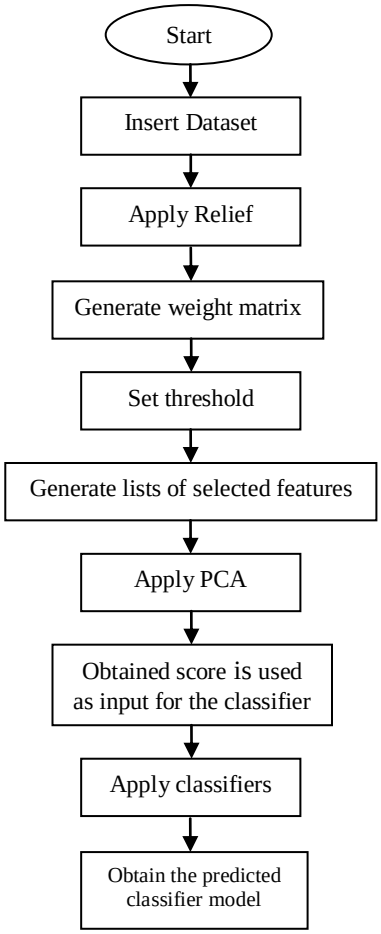

Figure 1 Proposed model

V. RESULTS AND DISCUSSION

It should include important findings discussed briefly. Wherever necessary, elaborate on the tables and figures without repeating their contents. Interpret the findings in view of the results obtained in this and in past studies on this topic. State the conclusions in a few sentences at the end of the paper. However, valid colored photographs can also be published.

V . EXPERIMENTAL RESULTS AND ILLUSTRATIONS

The results produced with the proposed cervical data set method are evaluated in this section. Analysing the dataset for any resulting analysis is the main objective of any experimental setup. The proposed methodology is evaluated on Cervical Dataset, which is taken from the UCI Repository. In terms of the Number of Selected Execution Time and Features, the work being proposed applies to cervical datasets and is evaluated.

Dataset Description

The cervical cancer dataset involves 858 samples and 32 features. Because the proposed method reduces the number of features to 21 and 18, with mean & median thresholds, A standard deviation threshold reduces the number of characteristics to just 2.

```
21 Selected Attributes
----------------------
1.  Age
2.  First sexual intercourse
3.  Num of pregnancies
4.  Smokes
5.  Smokes (years)
6.  Hormonal Contraceptives
7.  Hormonal Contraceptives (years)
8.  IUD
9.  IUD (years)
10.  STDs
11.  STDs (number)
12.  STDs:genital herpes
13.  STDs:HIV
14.  STDs: Number of diagnosis
15.  Dx:Cancer
16.  Dx:CIN
17.  Dx:HPV
18.  Dx
19.  Hinselmann
20.  Schiller
21.  Citology
----------------------
17 Selected Attributes
----------------------
1.  Age
2.  First sexual intercourse
```

Figure 2 Number of attributes selected on various threshold values

```
Results of Naive Bayes Classifier with 21 Attributes
------------------------------------------------------
Accuracy  : 86.512%
Precision : 100.000%
Recall    : 85.784%
F1 Score  : 92.348%
------------------------------------------------------
Results of Naive Bayes Classifier with 17 Attributes
------------------------------------------------------
Accuracy  : 92.558%
Precision : 99.476%
Recall    : 92.683%
F1 Score  : 95.960%
------------------------------------------------------
Results of Naive Bayes Classifier with 2 Attributes
------------------------------------------------------
Accuracy  : 93.023%
Precision : 98.936%
Recall    : 93.467%
F1 Score  : 96.124%
------------------------------------------------------
```

Figure 3 Naïve Bayes classifier

```
Results of SVM Classifier with 21 Attributes
--------------------------------------------
Accuracy  : 94.884%
Precision : 97.537%
Recall    : 97.059%
F1 Score  : 97.297%
--------------------------------------------
Results of SVM Classifier with 17 Attributes
--------------------------------------------
Accuracy  : 96.744%
Precision : 98.058%
Recall    : 98.537%
F1 Score  : 98.297%
--------------------------------------------
Results of SVM Classifier with 2 Attributes
--------------------------------------------
Accuracy  : 95.814%
Precision : 98.969%
Recall    : 96.482%
F1 Score  : 97.710%
--------------------------------------------
```

Figure 4 SVM classifier

```
Results of Random Forest Classifier with 21 Attributes
------------------------------------------------------
Accuracy  : 95.349%
Precision : 97.549%
Recall    : 97.549%
F1 Score  : 97.549%
------------------------------------------------------
Results of Random Forest Classifier with 17 Attributes
------------------------------------------------------
Accuracy  : 97.209%
Precision : 99.015%
Recall    : 98.049%
F1 Score  : 98.529%
------------------------------------------------------
Results of Random Forest Classifier with 2 Attributes
------------------------------------------------------
Accuracy  : 95.814%
Precision : 98.969%
Recall    : 96.482%
F1 Score  : 97.710%
------------------------------------------------------
```

Figure 5 Random Forest classifier

© 2020, IJCSE All Rights Reserved

Figure 1 visualizes the proposed model where all the steps have been mentioned for completing the implementation of the methodology.

Table 1 various accuracy parameters

Evaluation Criteria on varying threshold values		Classifiers		
		Naïve Bayes	SVM	Random Forest
Accuracy	21	86.51%	94.88%	95.34%
	17	92.55%	96.74%	97.20%
	2	93.02%	95.81%	95.81%
Precision	21	100%	97.53%	97.54%
	17	99.47%	98.05%	99.01%
	2	98.46%	98.81%	98.96%
Recall	21	85.78%	97.05%	97.54%
	17	92.68%	98.53%	98.04%
	2	93.46%	96.48%	96.43%
F1 measure	21	92.34%	97.29%	97.54%
	17	95.96%	98.29%	98.52%
	2	96.12%	97.71%	98.71%

Fig 6 Comparison of all the 3 algorithms on threshold value 21

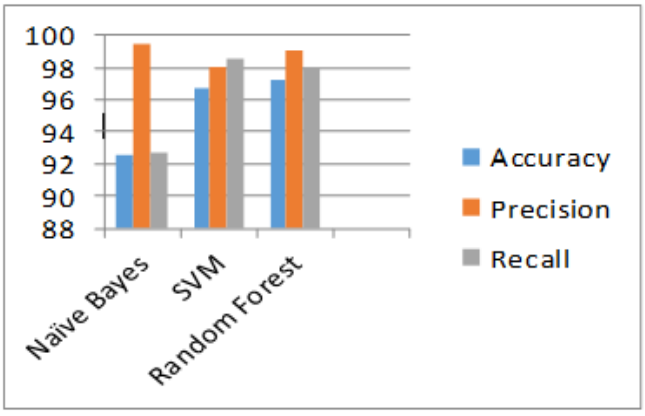

Figure 7 Comparison of all the 3 algorithms on threshold value 17

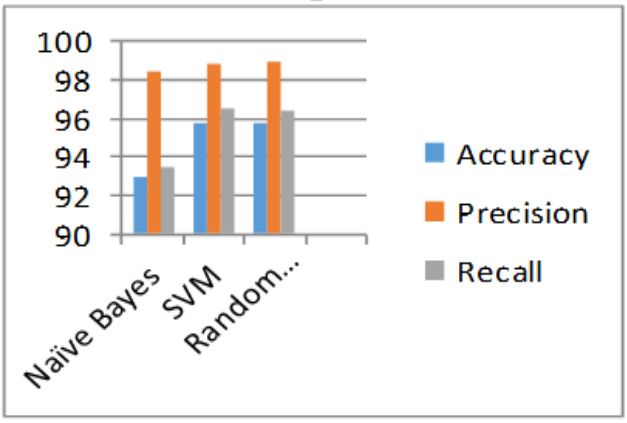

Figure 8 Comparison of all the 3 algorithms on threshold value 2

In Figures 6, 7 and 8, a comparison has been established which compares various accuracy parameters of Naïve Bayes, SVM and Random Forest. Among all these three algorithms Random Forest has shown the highest accuracy in all the three varying threshold values as compared to that of the other two algorithms.

VI. CONCLUSION

Cancer is probably a deadly disease. For medical physicians, cancer detection is still difficult. Until now it is not possible to invent the actual reason and cure cancer. Its key result of this study is to improve the doctor's survival diagnosis rate for cancer. It is achieved by predicting how many years of patients will survive to plan their next cancer therapy. A new Relief & PCA method based hybrid feature selection method is introduced. The method is tested on the data set for cervical cancer. Three classification algorithms namely, Naïve Bayes, SVM and Random Forest have been applied and it has been found that Random forest has the highest accuracy in all the three varying threshold values.

REFERENCES

[1] K.Arutchelvan, Dr.R.Periyasamy, "Cancer Prediction System Using Datamining Techniques" International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 08 | Nov-2015

[2] Ritu Chauhan "Data clustering method for Discovering clusters in spatial cancer databases" International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010. [

[3] Dechang Chen "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering" Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786.

[4] S M Halawani "A study of digital mammograms by using clustering algorithms" Journal of Scientific & Industrial Research Vol. 71, September 2012, pp. 594-600.

[5] Ada and Rajneet Kaur "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient" International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 4, April 2013, pg.1 – 6, ISSN 2320–088X

[6] V.Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646

© 2020, IJCSE All Rights Reserved

[7] Charles Edeki "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability" Mediterranean journal of Social Sciences Vol 3 (14) November 2012, ISSN: 2039-9340.
[8] Zakaria Suliman zubi "Improves Treatment Programs of Lung Cancer using Data Mining Techniques" Journal of Software Engineering and Applications, February 2014, 7, 69-77
[9] Labeed K Abdulgafoor "Detection of Brain Tumor using Modified K-Means Algorithm and SVM" International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013
[10] A. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org
[11] Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66.
[12] B Khalid, N Abdelwahab. â€œA Comparative Study of Various Data Mining Techniques: Statistics, Decision Trees and Neural Networksâ€•, International Journal of Computer Applications Technology and Research, Volume-5, Issue-03, pp (172 â€“ 175), 2016.
[13] S Mahajan, "Convergence of IT and Data Mining with other technologies ", International Journal of Scientific Research in Computer Science and Engineering, Volume-01, Issue-04, pp (31-37), Aug 2013

Authors Profile

Ms. Priyanka Rajpoot is a student of MITS Gwalior pursuing M. tech in Information and technology. I am currently doing my dissertation under the guidance of Mr. Mahesh Parmar (Assistant professor of CSE & IT) MITS Gwalior. I have earlier published my survey paper in **IJLTET- Volume 15 Issue 2, December 2019.**

Mahesh Parmar as an Assistant Professor in CSE&IT Department in MITS Gwalior and having 10 years of Academic and Professional experience. He received M.E. degree in Computer Engineering from SGSITS Indore. He has guided several students at Master and Under Graduate level. His areas of current research include Data mining and Image Processing. He has published more than 30 research papers in the journals and conferences of international repute. He has also published 02 book chapters. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, and IET etc.

© 2020, IJCSE All Rights Reserved