International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 1, Jan-Feb-2020, ISSN (Online): 2395-566X

Cancer Risk Prediction Models using Svm and Decision Tree

M. Tech. Scholar Priya Bharti **Prof. & H.O.D. Ashish Tiwari**
Priya.bhartibe17@gmail.com Ashishtiwari205@gmail.com
Dept. of Computer Science & Engineering
Vindhya Institute of Technology & Science
Indore, India

Abstract **- Cervical blight and the anticipation of analytic aftereffect are a part of the lot of important arising applications of gene announcement microarray technology with affection sequencing of microRNA. By application reliable and dependable allocation of apparatus acquirements algorithms accessible for microarray gene announcement profiling Data is the key in adjustment to advance the lot of acceptable and accessible predictive archetypal to be acclimated by patient. In this paper, two-machine acquirements algorithms accept been acclimated which are Support Vector Machine (SVM) and Decision Tree for the predictive models of cervical cancer. We analyze and appraise the achievement of these two algorithms in adjustment to apperceive which algorithm has bigger performance. In this study, 714 appearance and 58 samples are acclimated to advance predictive archetypal for cervical blight and our computational after-effects appearance that Decision Tree algorithm beat SVM algorithm with the accurateness of 99.42%. Our Data aswell accentuate the accent of variables, which accord the cogent role in admiration the accident of cervical cancer.**

Keywords- **Prevention, Cancer, Risk prediction.**

I. INTRODUCTION

Cervical blight and the anticipation of analytic aftereffect are a part of the a lot of important arising applications of gene announcement microarray technology with affection sequencing of microRNA. By application reliable and dependable allocation of apparatus acquirements algorithms accessible for microarray gene announcement profiling Data is the key in adjustment to advance the a lot of acceptable and accessible predictive archetypal to be acclimated by patient. In this paper, two-machine acquirements algorithms accept been acclimated which are Support Vector Machine (SVM) and Decision Tree for the predictive models of cervical cancer.

We analyze and appraise the achievement of these two algorithms in adjustment to apperceive which algorithm has bigger performance. In this study, 714 appearance and 58 samples are acclimated to advance predictive archetypal for cervical blight and our computational after-effects appearance that Decision Tree algorithm beat SVM algorithm with the accurateness of 99.42%. Our Data aswell accentuate the accent of variables, which accord the cogent role in admiration the accident of cervical cancer.

Cervical blight is the additional a lot of accepted blight a allotment of women. It may be able to cause bloodshed and anguish [1], [2]. The top bloodshed amount of the cervical blight is because of the abridgement of acquaintance of women for aboriginal apprehension of cervical blight [3]. Even admitting cervical blight is alarming which may advance to activity aggressive but it is potentially curable. Cervical blight occurs in a woman's cervix [4]. The cervix is the lower allotment of the uterus. It connects the uterus to the vagina [4]. There are two capital types of cervical cancer, which are squamous corpuscle blight and adenocarcinoma. The Squamous corpuscle blight blight blazon usually occurred in the epithelium lining of the cervix [5]. As for adenocarcinoma blight type, it is developed from gland cells. The adenocarcinoma of cervical blight is a premalignant glandular action [6]. Squamous corpuscle blight has the accomplished allotment of cervical blight cases which are 90% and 10% of cervical blight cases are adenocarcinoma [1].

Hence, adenocarcinoma neoplasia of the cervical blight is beneath accepted than squamous corpuscle blight neoplasia of cervical blight if it is getting diagnosed [6].Cervical blight is acquired by a virus called human papillomavirus (HPV). If the infection of the human papillomavirus (HPV) at the cervix larboard untreated, cervical blight is developed [7]. In cervical cancer, human papillomavirus (HPV) is the lot of important communicable abettor because it contributes to neoplastic progression.

Neoplastic progression is the progression of the aberrant advance of the cervical annihilative beef and admeasurement of the aberrant beef due to a annihilative

action [8]. However, a lot of accurate studies accept begin that human papillomavirus (HPV) infection abandoned is bare to abet the annihilative of cervical cancer. Other host abiogenetic variations aswell play important role in the development of cervical blight in women [1]. Commonly, microarray gene announcement profiling is acclimated to assay and differentiate the bidding gene announcement in a precancerous and annihilative corpuscle in the cervix [1]. A microarray or aswell accepted as DNA dent is an adjustment of DNA molecules that accept been chemically affirmed to a accomplished filigree of surfaces.

The purpose of the microarray gene announcement is to adapt and assay the genes announcement accompaniment in commutual DNA able from mRNA in which the admixture is demography abode on the arrangement [9]–[11]. Genome-wide announcement profiling has the abeyant to get added accurately adumbrate aftereffect at the end of the analysis back it allows alternative of the genes that accessory a lot of acerb with the aftereffect by screening a ample amount of these genes simultaneously. Announcement profiling studies appearance that genes are statistically cogent differences beneath afflicted beginning altitude [12]. In this study, the predictive archetypal for free whether the beef are annihilative or precancerous for cervical blight can be developed by just searching at their gene announcement profiles. In adjustment to advance the predictive model, apparatus acquirements algorithms approaches accept been implemented.

Artificial intelligence and apparatus acquirements techniques accept been activated in assorted medical analytic systems [13]–[16] and there are abounding advisers activated the techniques in the angel processing of Pap apply images [17]–[23]. However, there are beneath studies on implementing apparatus acquirements methods on the cervical blight based on gene announcement profiling data. Hence, in adjustment to affected this problem, we implemented unsupervised and supervised apparatus acquirements techniques, which are Principal Components Analysis (PCA), Support Vector Machine (SVM) and Random Forest (DT), respectively.

The Principal Components Analysis (PCA) is acclimated to do the Data screening and Data preprocessing for the model. The Data that accept been background in the Principal Components Analysis (PCA) will be acclimated in Support Vector Machine (SVM). As for abutment apparatus agent (SVM) and Random Forest (DT), they are acclimated to adumbrate the accurateness of the predictive archetypal for cervical blight whether the corpuscle is pre-cancerous or annihilative cells.

II. PROPOSED METHODS

In this study, we acclimated dataset, which is acquired from the Gynecologic Oncology Group Tissue Bank (PA, USA) for allocation of cervical blight based on the gene announcement profiling Data [24]. First, the dataset will be ability Data abstraction action in adjustment to get and retrieving the accordant Data or advice in the dataset. The Data pre-processing again activated on the dataset for eliminating the extraneous and bombastic Data absolute in the dataset. In adjustment to accomplish allocation of the gene announcement profiling Data of microarray dataset into their cluster, a tree-like anatomy is complete by application hierarchical clustering. By implemented apparatus acquirements algorithms of Support Vector Machine (SVM) and Decision Tree (DT) on the dataset, predictive archetypal for cervical blight can be developed.

In Support Vector Machine (SVM) model, Principal Components Analysis (PCA) algorithm is acclimated as for the affection alternative technique. It is actual advantageous in adjustment to abate the top ambit of dataset as to facilitate the algorithm to aftermath a acceptable performance. The new dataset complete will be accomplished by application Support Vector Machine (SVM) archetypal in adjustment to get the anticipation accuracy. As a comparison, Decision Tree (DT) is aswell activated to get the accurateness of predictive archetypal for cervical cancer.

1. Data Abstraction

Data abstraction is a action of retrieving accordant Data or information out of Data sources, which is dataset absolute microarray Data analysis. In this stage, the Data will be extracted and loaded into the staging breadth of the relational database of the cervical cancer. When Data abstraction is implemented, the abstraction argumentation is activated in adjustment to abstract the Data from the dataset. The Data cavalcade of the dataset is the Data apropos the normal and tumor of samples data. And as for the Data in the row is about the appearance of the sequencing of microRNAs. The Data cast aftermath from the Data abstraction is afflicted into a Data anatomy because it may affluence us to adapt the Data for added processing.

2. DataPre-Processing

Data pre-processing is an important footfall in the Data mining process. It is implemented on the dataset because the dataset acquired from microarray assay contains irrelevant, capricious and bombastic of Data or babble present in the dataset. Data pre-processing is performed by ability the action of Data cleaning, Data ascent and normalization of data. In Data charwoman process, the base or extraneous Data from the datasets are detected. Then, the Data that accept been articular as abridged and extraneous will be filtered out of the dataset. After charwoman the besmirched data, a final dataset should be constant with added agnate dataset in the arrangement and

can be acclimated for added processing. Added than Data charwoman process, Data ascent is aswell implemented on the dataset to pre-processing the dataset. In Data scaling, the Data will be clustering and ascent whenever there are Data with altered units and ranges. By ascent the data, the final dataset will be simple to be interpreted. The beggarly amount of sample Data from the datasets is bargain by application the clustering transformation method. As for ascent transformation, it will bisect the amount of anniversary sample Data by the accepted aberration of the dataset. Thus, the after Data of the dataset adherence is improved. The normalization of the Data is important to adapt the columns and tables of the database to abate the back-up of the data. It is acclimated to advance the candor of the dataset in this study.

3. Hierarchical Clustering

The subgroups are authentic based on the genes announcement profiling Data of the dataset. In adjustment to allocate the gene announcement profiles of Data into their own clusters, hierarchical clustering is implemented to actualize the hierarchical, which is the tree-like anatomy of the data. It is aswell referred as a dendrogram. The hierarchical clustering is implemented by barometer the affection similarities of the gene announcement profiles. Hierarchical clustering is implemented by barometer the best ambit cast amid the samples data. In every date of hierarchical clustering, the two abutting clusters are alloyed calm into a new array or subtypes. The clustering action is again until all of the samples Data is agglomerated into an individual cluster.

4. Principal Components Analysis (PCA)

In adjustment to advance the predictive archetypal by application the Support Vector Machine (SVM), Principal Components Analysis (PCA) is acclimated as dimensional abridgement address because it is advantageous to collapse the appearance into a abate set of arch components. The Principal Components Analysis (PCA) is acclimated in this abstraction because the dataset contains hundreds of appearance from the sequencing of microRNA. The Principal Components Analysis (PCA) is bare in adjustment to abate the ambit of the hundreds of appearance in the dataset. By implemented the apparatus acquirements algorithm on the dataset, it helps to collapse the hundreds of appearance into a abate set of arch components.

5. Support Vector Machine (SVM)

Support Vector Machine (SVM) algorithm is implemented by assuming an allocation on the final dataset. The allocation of the algorithm is implemented by amalgam a multidimensional hyperplane to the careful samples data. It will aerate the allowance amid the two-data clusters, which are normal (N) and tumor (T). The blazon of the SVM constant acclimated in this abstraction is set to the Cclassification. As for the kernel, beeline of atom blazon is selected. In adjustment to appraise the predictive archetypal performance, allocation of accuracy, acuteness and specificity charge to be calculated, as apparent in (1), (2) and (3).

Accuracy = (TP+TN)/(TP+FP+FN+TN) (1)

Sensitivity= TN/(TP+FN) (2)

Specificity= TP/(FP+TP) (3)

TN is the number of true negative, FP is defined as the number of false positive, TP is defined as number of true positive and FN is the number of false negative. These coefficients are defined and illustrated as a confusion matrix in the Table 1

Table I: Confusion matrix.

		Reference	
		Negative (0)	Positive (1)
Prediction	Negative (0)	TN	FP
	Positive (1)	FN	TP

III. RESULTS AND DISCUSSION

1. Analysis:

For the analysis we have taken following data for the result: In our research, the dataset encounters the class imbalance problem. Out of 3600 patients, 2880 patients were satisfied with control of cancer, which constitutes about 80 % of the total patients and 720 patients are unsatisfied. The imbalanced ratio equals 4:1 between majority and minority. In other words, a dataset is class-imbalanced if one class includes significantly more sample numbers than the other. In order to resolve the problem, we can pick the random under sampling (RUS), random over sampling (ROS), and Synthetic Minority Oversampling Technique (SMOTE), which are among the most used resampling methods to counterpoise imbalanced datasets. Here, we only choose SMOTE algorithms, which are used to create one more dataset, where the minority samples were oversampled by 400% and the majority class was under sampled at 1% to approximately make the ratio 1:1. The descriptions of the datasets are given in Table 3.1. Eventually, the balanced dataset was used to construct the model.

Table II: Cervical cancer data set description before attribute reduction.

Attributes	Types	Attributes2	Types3
Age	Int	STDs (number)	Int
Number of sexual partnes	Int	STDs: condylomatosis	Bool
Age of first sexual Intercourse	Int	STDs: cervical condylomatosis	Bool
Number of pregnancies	Int	STDs: vaginal condylomatosis	Bool
Smokes	Bool	STDs:vulvo-perineal condylomatosi	Bool
Smokes (years)	Bool	STDs: syphilis	Bool
Smokes (packs/year)	Bool	STDs: pelvic inflammatory disease	Bool
Hormonal Contraceptives	Bool	STDs: genital herpes	Bool
Hormonal Contraceptives (years)	Int	STDs: molluscum contagiosum	Bool
Intra Uterine Device (IUD)	Bool	STDs: AIDS	Bool
IUD (years)	Int	STDs: HIV	Bool
Sexually Transmitted Diseases (STDs)	Bool	STDs: Hepatitis B	Bool

2. Result Compression of Our Research Methodology:

As can be seen from Table 3.2 and graph 3.1, in this study, the performance of the four final predictive models was evaluated using G-mean. For the testing dataset, the final comparative analysis results demonstrated that the Decision Tree algorithm showed the best with accuracy of 99.42%, and the sensitivity and specificity were 91.89% and 86.24%, respectively. The SVM algorithm came out to be the second best with a classification accuracy of 95.06%, and the sensitivity and specificity gave 98.06% and 86.05%, respectively. The Adaboost algorithm came out to be the third best with a classification accuracy of 94.77%, and the sensitivity and specificity gave 98.83% and 82.56%, respectively. The Bagging algorithm came out to be the fourth best with a classification accuracy of 91.86%, and the sensitivity and specificity gave 98.84% and 70.93%, respectively. In the results, the area under the G-values of the SVM, Adaboost, Bagging, and decision tree algorithms were 91.86, 90.32, 83.73, and 89.02, respectively.

3. Result Compression of Our Research Methodology

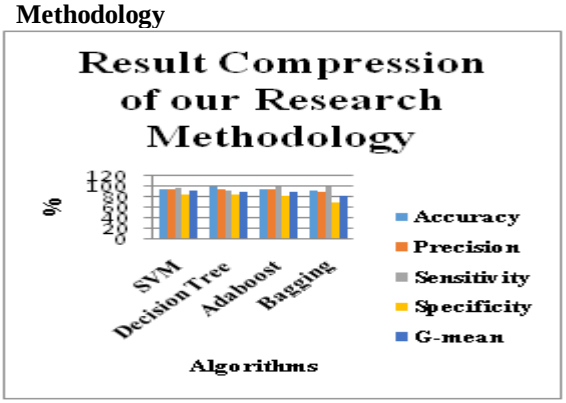

Fig.1 Result Compression of our Research Methodology.

IV. CONCLUSION

In this study, we used dataset, which is obtained from the Gynecologic Oncology Group Tissue Bank (PA, USA) [24] for classification of cervical cancer based on the gene expression profiling data. We conducted the comprehensive benchmarking of Support Vector Machine (SVM) and Decision Tree to evaluate the performance of the predictive models based on the accuracy, Kappa value, sensitivity and specificity. Based on our computational result, we can conclude and prove that Decision Tree machine learning algorithm can be successfully used for Predicting cervical cancer based on the gene expression profiling data with the microarray dataset. The accuracy obtained in cervical cancer prediction is better than previous researches results. The average Decision Tree model's accuracy obtained is 99.42%, which may be acceptable in many applications. Besides that, the investigation for a high overall performance classifier with microarray gene expression and other "omics" data is still in progress. Decision Tree algorithms exhibit one of the best classification performances in predicting cervical cancer.

REFERENCE

[1]. C. M. Martin, K. Astbury, L. McEvoy, S. O'Toole, O. Sheils, and J. J. O'Leary, "Gene expression profiling in cervical cancer: identification of novel markers for disease diagnosis and therapy.," Methods Mol. Biol., vol. 511, pp. 333–359, 2009.

[2]. M. Lupu, C. Caruntu, M. A. Ghita, V. Voiculescu, S. Voiculescu, A. E. Rosca, A. Caruntu, L. Moraru, I. M. Popa, B. Calenic, M. Greabu, and D. E. Costea, "Gene Expression and Proteome Analysis as Sources of Biomarkers in Basal Cell Carcinoma," Disease Markers, vol. 2016. 2016.

[3]. S. W. Purnami, P. M. Khasanah, S. H. Sumartini, V. Chosuvivatwong, and H. Sriplung, "Cervical cancer survival prediction using hybrid of SMOTE, CART and smooth support vector machine," in AIP Conference Proceedings, 2016, vol. 1723.

[4]. D. R. Tobergte and S. Curtis, Principles of Biochemistry 5th ed, vol. 53, no. 9. 2013.

[5]. M. Volpacchio, J. C. Vilanova, and A. Luna, "Cervix and vagina," Learn. Imaging, vol. 1, no. 6, pp. 165–186, 2012.

[6]. M. P. Hopkins, J. a Roberts, and R. W. Schmidt, "Cervical adenocarcinoma in situ.," Obstetrics and gynecology, vol. 71, no. 6 Pt 1. pp. 842–844, 1988.

[7]. A. Zayed, "Cervical Cancer: Types, Symptoms, Causes, Treatments and Medications,"https://www.consumerhealthdigest.com/health-conditions/cervical-cancer.html.

[8]. A. Deverakonda and N. Gupta, "Diagnosis and treatment of cervical cancer: a review," J. Nurs.Heal. Sci., vol. 2, no. 3, pp. 1–11, 2016.

[9]. "DNA Microarray microarray," in Encyclopedia of Public Health, 2008, pp. 298–299.

[10]. M. Debnath, G. B. K. S. Prasad, and P. S. Bisen, "Microarray," in Molecular Diagnostics: Promises and Possibilities, 2010, pp. 193–208.

[11].A. M. Charpe, "DNA microarray," in Advances in Biotechnology, vol. 9788132215547, 2013, pp. 71–104.
[12].A. H. Klopp and P. J. Eifel, "Gene expression profiling in cervical cancer: State of the art and future directions," Cancer Journal, vol. 12, no. 3. pp. 170–174, 2006.
[13].A. A. Abdullah, B. S. Chize, and Y. Nishio, "Implementation of an improved cellular neural network algorithm for brain tumor detection," in 2012 International Conference on BiomedicalEngineering (ICoBE), 2012, pp. 611–615.
[14].A. A. Abdullah and S. Kanaya, "Prediction of Biological Activities of Volatile Metabolites Using Molecular Fingerprints and Machine Learning Methods," J. Telecommun. Electron.Comput. Eng., vol. 10, no. 1–17, pp. 91–96, 2018.
[15]. A. A. Abdullah, N. S. Fadil, and W. Khairunizam, "Development of Fuzzy Expert System for Diagnosis of Diabetes," in 2018 International Conference on Computational Approach inSmart Systems Design and Applications, ICASSDA 2018, 2018.
[16]. A. A. Abdullah, A. Yaakob, and Z. Ibrahim, "Prediction of Spinal Abnormalities Using Machine Learning Techniques," in 2018 International Conference on Computational Approachin Smart Systems Design and Applications, ICASSDA 2018, 2018.
[17]. A. A. Abdullah, A. Fonetta, D. Giong, N. Adilah, and H. Zahri, "Cervical cancer detection method using an improved cellular neural network (CNN) algorithm," Indones. J. Electr. Eng.Comput. Sci., vol. 14, no. 1, pp. 210–218, 2019.
[18]. N. A. Obukhova, A. A. Motyko, U. Kang, S.-J. Bae, and D.-S. Lee, "Automated image analysis in multispectral system for cervical cancer diagnostic," Conf. Open Innov. Assoc. Fruct, vol. 2017–April, 2017.
[19]. P. Sukumar and R. K. Gnanamurthy, "Computer Aided Detection of Cervical Cancer Using Pap Smear Images Based on Adaptive Neuro Fuzzy Inference System Classifier," J. Med.Imaging Heal. Informatics, vol. 6, no. 2, pp. 312–319, 2016.
[20]. W. William, A. Ware, A. H. Basaza-Ejiri, and J. Obungoloch, "A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images," Computer Methods and Programs in Biomedicine, vol. 164. pp. 15–22, 2018.M. Orozco-Monteagudo, A. Taboada-Crispi, and H. Sahli, "Biologically inspired anomaly detection in pap-smear images," in Lecture Notes in Computer Science (including subseriesLecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, vol. 8259LNCS, no. PART 2, pp. 17–24.
[21]. B. Taha, J. Dias, and N. Werghi, "Classification of cervical-cancer using pap-smear images: A convolutional neural network approach," in Communications in Computer and InformationScience, 2017, vol. 723, pp. 261–272.
[22]. Abid Sarwar, "Analysis of Machine Learning and Statistics Tool Box (Matlab R2016) over Novel Benchmark Cervical Cancer Database," Int. J. Trend Sci. Res. Dev., vol. 2, no. 1, pp. 619–622, 2018.
[23]. D. Witten, R. Tibshirani, S. G. Gu, A. Fire, and W. O. Lui, "Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls," BMC Biol., vol. 8, 2010.
[24]. Z. Yue, Z. Yun-Shan, and X. Feng-Xia, "miR-205 mediates the inhibition of cervical cancer cell proliferation using olmesartan," JRAAS - J. Renin-Angiotensin-Aldosterone Syst., vol. 17, no. 3, 2016.
[25]. H. Xie, Y. Zhao, S. Caramuta, C. Larsson, and W. O. Lui, "miR-205 Expression Promotes Cell Proliferation and Migration of Human Cervical Cancer Cells," PLoS One, vol. 7, no. 10, 2012.
[26]. R. Hou, D. Wang, and J. Lu, "MicroRNA-10b inhibits proliferation, migration and invasion in cervical cancer cells via direct targeting of insulin-like growth factor-1 receptor," Oncol. Lett., vol. 13, no. 6, pp. 5009–5015, 2017.
[27]. D. Zou, Q. Zhou, D. Wang, L. Guan, L. Yuan, and S. Li, "The Downregulation of MicroRNA-10b and its Role in Cervical Cancer," Oncol. Res. Featur. Preclin. Clin. Cancer Ther., vol. 24, no. 2, pp. 99–108, 2016.
[28]. B. Pardini, D. De Maria, A. Francavilla, C. Di Gaetano, G. Ronco, and A. Naccarati, "MicroRNAs as markers of progression in cervical cancer: A systematic review," BMCCancer, vol. 18, no. 1, 2018.