ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse121942020.pdf
https://doi.org/10.30534/ijatcse/2020/121942020

A WCO Based Cancer Survival Prediction Using Statistical Feature Selection

[1] Sanku Rajendra Kumar, [2]Dr. Anil Kumar
[1] Research Scholar, Dept. of Computer Science and Engineering, Sri Satya Sai University & Medical Science, Sehore, Madhya Pradesh, India.
Email: rajendhar.sanku@gmail.com
[2] Associate Professor, Dept. of Computer Science and Engineering, Sri Satya Sai University & Medical Science, Sehore, Madhya Pradesh, India.

ABSTRACT

The heavily perceived illness is the cancer and it is answerable for wide range of deaths consistently. In spite of the way that cancer is remediable and curable in most early phases the most patients are analyzed with cancer extremely delay. The data mining procedure and characterization are a proficient method to arrange the information especially in clinical fields, where those methodologies are comprehensively utilized in conclusion to settle on choice. The wisconsin cancer original (WCO) dataset is utilized. The feature selection assumes a significant job in cancer order, for quality articulation information as a rule have countless measurements and generally few examples. Gene selection is a mainstream innovation for cancer grouping that intends to recognize few useful qualities from a large number of qualities that may add to the event of diseases to acquire a high predictive precision. A specific feature selection technique is run on various subsamples and the derived features are intersected into a steadier subset. The prediction of survival result, for example, infection explicit or generally survival after disease diagnosis or treatment is the principle objective. The prediction of cancer outcome usually refers to the cases of longevity, perseverance, movement and surgery susceptibility. This prediction will be worked on the cancer information accessible from the SEER (surveillance epidemiology and end results) technique with point of creating precise survival expectation models for cancer. So in this paper, a WCO based cancer prediction using statistical feature selection will be implemented.

Key words : Prediction, Cancer, data mining, wisconsin cancer original (WCO), SEER, feature selection, survival.

1.INTRODUCTION

The colon and rectum cancers are two major kinds of cancers available in the world. Early diagnosis and treatment can enormously improve the odds of survivability [1]. The SEER technique is a valuable source of household measurements of cancer disease. The gathered information by the SEER speaks to most of the US populace over a few geographic districts. This information is derivable from the SEER site after presenting a SEER restricted use information understanding structure. In this paper, study the colon disease information accessible from the SEER method with the point of creating precise existence expectation models for cancer. The information broke down in this investigation is from the observation, the study of SEER colon and rectum cancer growth occurrence information in the long periods. The SEER colon and rectum disease frequency information comprise of four information sets. Here we utilize administered grouping techniques to anticipate existence of cancer growth patients, toward the finish of diagnosis on year basis. It perform different modes of observations to determine that several meta classifiers utilized with choice trees and capacities will be able to provide perfect outcomes contrasted with fundamental classifiers [2]. These outcomes can be improved by receiving SMOTE (Synthetic Minority Over-sampling Technique) to balance the existence and non-existence classes, and by consolidating the subsequent forecast probabilities from a few classifiers utilizing a group casting a ballot scheme.

The cancer will be developing to around 20 million by 2025. Presenting suitable devices for recognizing health professional's cancer cases can have a significant advance in therapy conclusion of patients of cancer. This disease has lot of illnesses such as colon cancer, breast cancer and lung cancer. A wide range of cancer, a portion of the body's cells starts to separate ceaselessly and spread into surrounding tissues. This disease may begin from any place in the body trillions of cells will be formed from it. Regularly, human cells increment and cut into new depending on requirement. The injured cells and olden cells are expired and after this new form of cells formed in place of them. At the time creating this disease is the primary procedure is separates and old harmed cells are made due during the time at cells gone expired [5]. At the point when it isn't required new cells are expands at that point. Tumors are developed by additional cells expanding hence this procedure is progressing approach. The cancer cells are expanded randomly, and it is not the same as some other ordinary cells. The basic

difference among the typical cells and disease cells is that ordinary cells have large scope to get cancer. This disease is a very dangerous and serious sickness and it can cause a change in the body of a human and gradually these cells are increased and get into partitioned in the human body. Due to this the resistance power of a human body will have chance to diminish to a max extent. The metastatic cancer is the disease which is arises somewhere in the body of a human and spread those cells to entire body gradually. The term metastasis is nothing but the way how the cancer cells are distributed in various locations of a human body.

The concept of data mining is expanding quickly in the clinical field because of its achievement in the characterization and forecast algorithms which assists specialists in dynamic. This technique is intended to searching for approaches to enhance medical condition of a patient and diminish expenses of medication; data mining causes a great deal to satisfy this reason. Data mining is currently extremely well known. Anticipating illness is currently likewise an extraordinary test [4]. This technique is capable to predict the mode of illness. It once in a while found the paper which partner cancer illness expectation. The cancer is very harmful illness and is developing quickly. The obscure is the one kind of cancer. Even experts can't have a clue about the best possible reasons for cancer. The people are worried more about their future if they have been attacked with a dangerous disease called cancer. Most of the people in the world are planned to identify the cancer disease in a very initial stage with help of the past knowledge based on symptoms. So in this way people will get idea about the disease and hence they will take necessary prevention steps and be conscious about the disease. There is a analysis for knowing information about cancer with WEKA (waikato environment for knowledge analysis) which is a open source for data mining technique. The technique is much useful to estimate the sickness of cancer with a great accuracy and so that it can assist the people in a view of taking required action perfectly. The cancer diseases are estimated with more accuracy with help of data mining tool. This tool can use three famous algorithms such as Naive Bayes, K-Nearest and j48 algorithm. These algorithms utilize a dataset for the purpose of providing better solution. This dataset can be verified thoroughly by the expert and intelligent team and they can decide the performance of various algorithms with respect to their accuracy. These algorithms can also provide complete information about cancer patients.

2.LITERATURE SURVEY

The reason for this research is to give a review of literature through a diagram of the examination fields pertinent to cancer treatment and expectation approaches dependent on AI (Artificial Intelligence). The previous hardly any years have seen an exponential development in data bases and repositories because of the expansion in logical information and the generation of huge amount of information [3]. Biomedical domain speaks to one of the rich information areas. A broad measure of biomedical information is as of now accessible with abundance of information, extending from complete details of clinical side effects to different sorts of biochemical information and output of imaging device. One of the significant biomedical exploration areas is epidemiological cancer research, who is of high need over the world. This cancer has been portrayed as a heterogeneous infection comprising of a wide range of subtypes. Each 6th demise on the planet is because of cancer, making it the subsequent driving reason for death. The early determination and anticipation of a disease type have become a need in cancer research, as it can encourage the resulting clinical administration of patients. The immense measure of hidden information in gigantic data bases identified with cancer with all his inconstancy has made colossal interests in the field of data mining [6]. While Data mining is an order coming about because of the blend of old style insights and software engineering algorithms, for example, Machine Learning, intend to the extraction of new and valuable hidden information from a lot of information, it has become a helpful instrument in Bioinformatics.

It can portray variety of cancer frequency and mortality by locale, ethnicity, sexual orientation and financial components that add to the evaluation of populace health needs, while it can add to the investigation of disease trouble. Besides, top to bottom examination of the patient's profile utilizing data mining strategies may reveal covered up, already obscure relations between understanding profile, cancer treatment and observation. This section will introduce late 325 methodologies and works identified with this specific circumstance. It presents synthesis of current and future patterns for Smart Systems for E-health. In late decades, due to the drastically and exponential development of cancer occurrence and related deaths around the world, critical advancement has been made in the improvement of data mining. The various data mining procedures and approaches are applied to study and summarize information gained from disease data bases, so it can derive significant information [7]. The survey of various data mining applications is studied in the region of cancer conclusion and expectation. The early analysis and anticipation of a cancer make it bound to react to successful treatment and can bring about a more probability of existence, less bleakness, and more affordable treatment.

Huge enhancements can be made in the lives of cancer patients by distinguishing disease early and maintaining a strategic distance from delays in care. The measure of information originating from medical investigation of this illness is very enormous. This information assembles an immense and rich dataset [8]. The artificial intelligence (AI) in medical field has became very popular after it starts to applied in machine learning techniques to predict the cancer data bases and, with inside and out examination of the patient's profile utilizing statistical or data mining techniques may reveal covered up, already obscure relations between

patient profile, cancer treatment and observation and Consequently help resolving diagnostic and prognostic issues identified with cancer. So in this, the prediction of the cancer using statistical selection feature will be determined. The cancer survival prediction using statistical feature selection (CSP-SFS) is implemented for Wisconsin cancer original.

3.THE CSP-SFS CLASSIFIER

The data bases are amazingly powerless against missing, distortion and uncertain information as a result of their immense size and are begun from different incidental sources. Accordingly, pre-preparing is significant stage which is trailed by classification. The middle of the individual attributes values are most useful in the concept of pre-processing of dataset and this can able to determine the missing values by using that middle value. There is one data set called WCO dataset and it has a property atomic cell which consists of 17 missing qualities. In this way it can be easy to preprocess attributes physically. All the missing values are replaced with help of middle. So in this way dataset will be free from missing qualities. There are four data mining approaches such as Bayes classifier (Naive Bayes, Bayesian Logistic Regression), Decision Tree (J48, simple CART). All these classifiers are much useful in estimating dataset of cancer. By considering all these classifiers and WEKA apparatus, it is possible to discover what of this classifier will show the most exact outcome for cancer.

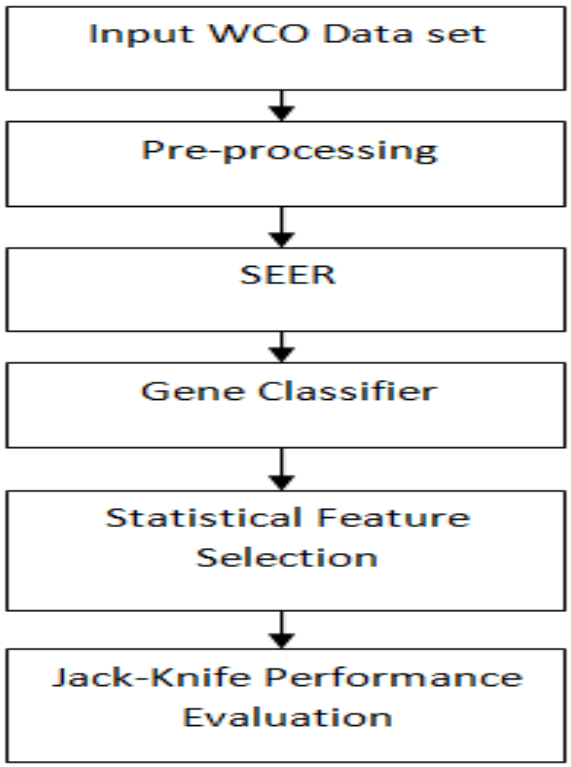

Figure. 1: Block diagram of CSP-SFS system

Data mining devices were utilized for the forecast of cancer, principle reason for existing was to group Naive Bayes (NB) algorithm, Bayesian Logistic Regression, Simple CART and J48 based on certain specifications. Dataset was gathered from wisconsin cancer original (WCO). The decision Tree algorithm was utilized to part entire information into subsets while J48 depended on choice hub and these hubs predict the normal information from entire arrangement of information. The SEER data base was utilized to describe the bosom malignant growth survivability rate. The SEER information is more dependable for expectation of various phases of bosom malignant growth. The SEER Program gives information on malignant growth measurements with an end goal to decrease the disease trouble. With SEER information being freely accessible, there is an experienced writing on SEER information contemplates. The SEER gives SEER*Stat a measurable programming which gives accommodation to analyze the information. What's more, there have been data mining applications produced for different kinds of malignant growth dependent on SEER information. Various methods dependent on data mining have been proposed for the survivability examination of different sorts of disease related to cancer. The classification methods utilized in our tests are of two kinds such as basic classifiers, and meta classifiers. The basic classifiers comprise of trees, tasks, and statistical techniques. The support for the classifiers in view of enhance their execution will be achieved by aid of meta classifiers.

Cancer classification based on gene articulation information contains countless features, which needs a generally huge preparing set to become familiar with a classifier with a low mistake rate. The FS and feature extraction methods on Artificial Neural Network (ANN), Support Vector Machine (SVM) and Naive Bayes (NB) were applied for the expectation of bosom malignant growth. Dataset of patients was gathered from wisconsin diagnostic cancer. The feature selection is a determination of sub features from an enormous dataset that helps in algorithm process. Creators nearly investigated every method with various kind of feature selection, for example, coloration based feature selection (CFS), and linear discriminated Analysis and Recursive Feature Elimination (RFE). After the relative investigation with various feature selection techniques, creators came to realize that the precision pace of Artificial Neural Network was higher than different algorithms.

The prediction of the cancer based on machine learning technique will be implemented based on devices, information sources, information type, information pre-preparing technique, information assessment strategy, approval technique and precision level of every algorithm in various circumstances. The relative information gain for every one of these 13 attributes. The information is introduced next to each other for the three times of enthusiasm alongside the normal. A jack-knife estimator is have computer estimate forgetting about one observation thus from the expectation test. The n expectations grant the tendency and change of the estimation

to be resolved and exactly when the information are not self-governing cross-evaluation ends up being more problematic as overlooking an observation doesn't empty all the related information in view of the associations with various recognitions. In this way this technique is an effective method to predict the cancer based on WCO with help of feature selection method.

4.RESULTS

The percentage of accuracy for various algorithms will be show in table 1. From this table it is clear that the present CSP-SFS classifier was given better percentage of accuracy with 98.8519. That means it can enhance the accuracy of a prediction of cancer using feature selection method. The accuracy is very less in bayesian logistic regression (BLR) compared to the remaining all. The figure. 2 represents the accuracy of different classifiers. The time that required for execution for various algorithms will be show in table 2. From this table it is clear that the present CSP-SFS classifier was executed with least time around 0.018 seconds only. That means it can enhance the execution time of a prediction of cancer using feature selection method. The figure. 3 represents the execution time of different classifiers. The precision for various classifiers will be show in table 3. From this table it is clear that the present CSP-SFS classifier has produced better precision around 97.3. That means it can enhance the precision of a prediction of cancer using feature selection method. So the missing attributes of dataset can easily predicted using this present classifier. The figure. 4 represents the precision of different classifiers. So by considering all the parameters such as accuracy, precision and execution time for various classifiers, it can conclude that the present classifier called CSP-SFS is have better accuracy, execution time and precision. This can be made possible by using cancer prediction with help of statistical feature selection method.

Table 1: The accuracy of WCO Dataset for various classifiers

S No	Classifier	Percentage of Accuracy
1	CSP-SFS	98.8519
2	Naive Bayes	94.6254
3	BLR	66.7853
4	Simple CART	97.9247
5	J48	97.5639

Table 2: Execution time for various classifiers.

S No	Algorithm	Time require for execution(s)
1	CSP-SFS	0.018
2	Naive Bayes	0.031
3	BLR	0.057
4	Simple CART	0.294
5	J48	0.236

Table 3:. Precision for various classifiers

S No	Classifier	Precision
1	CSP-SFS	97.3
2	J48	94.6
3	Naive Bayes	92.7
4	BLR	91.3
5	Simple CART	90.4

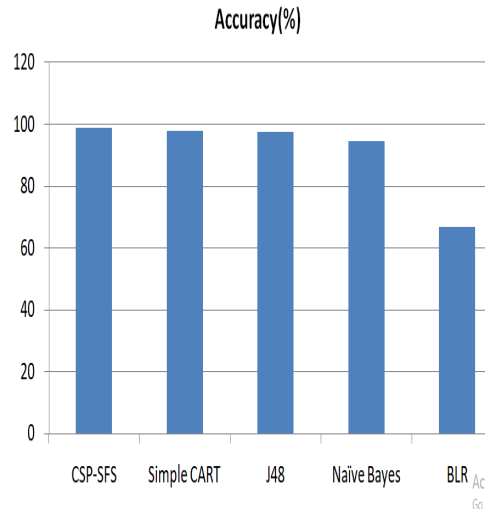

Figure 2: Percentage of accuracy for different classifiers.

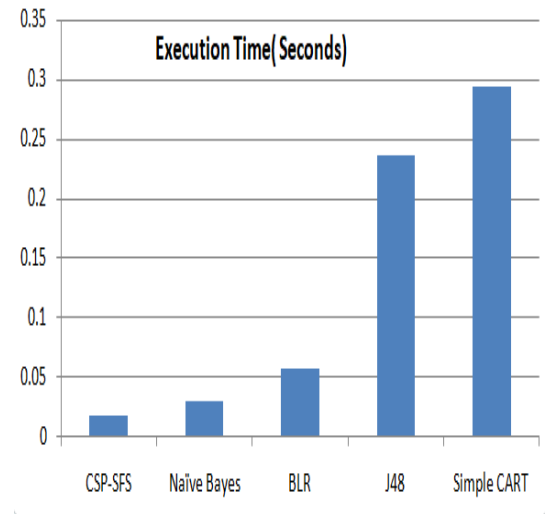

Figure 3: The execution time for different classifiers.

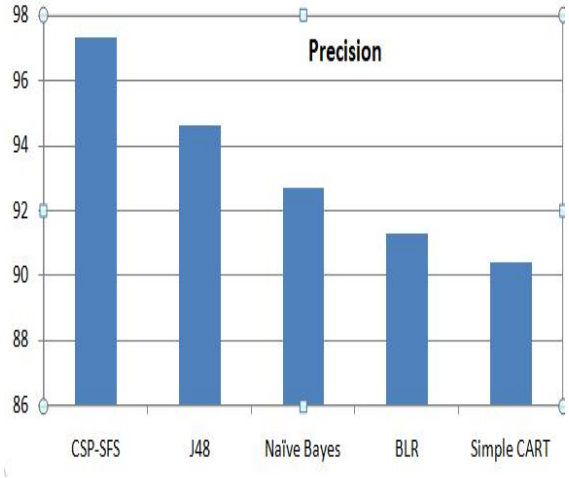

Figure. 4: The Precision for Different Classifiers.

5.CONCLUSION

The wisconsin cancer original (WCO) dataset was utilized to predict the cancer. The feature selection performed a significant job in cancer disease, for quality articulation information as a rule have countless measurements and generally few examples. Gene selection was a mainstream innovation for cancer grouping that intends to recognize few useful qualities from a large number of qualities that may add to the event of diseases to acquire a high predictive precision. The cancer is remediable and curable in most early phases the most patients are analyzed with cancer extremely delay. The prediction of cancer outcome usually refers to the cases of longevity, perseverance, movement and surgery susceptibility. A specific feature selection (FS) technique is run on various subsamples and the derived features are intersected into a steadier subset. The FS has worked on the cancer information accessible from the SEER (surveillance epidemiology and end results) technique with point of creating precise survival expectation models for cancer. So in this paper, a WCO based cancer prediction using statistical feature selection has been implemented.

REFERENCES

[1] Enas M.F. El Houby (2018**) A survey on applying machine learning techniques for management of diseases**, Journal of Applied Biomedicine https://doi.org/10.1016/j.jab.2018.01.002

[2] Dona Sara Jacob, Rakhi Viswan, V Manju, L PadmaSuresh, Shine Raj, "**A Survey on Breast Cancer Prediction Using Data mining Techniques**", IEEE Access, 2018, ISBN: 978-1-5386-3479-0.

[3] B.M.S.Rani, Dr .A .Jhansi Rani **,"Biometric Retinal Security System for User Identification and Authentication using GUI,**" in PONTE International Journal of Science and Research Vol .74, No.3/1, Mar 2018.(SCI)

[4] B.M.S.Rani and Dr.A.JhansiRani,"**A Hybrid Biometric Identification And Authentication System With Retinal Verification Using AWN Classifier For Enhancing Security**", AICC-2018 SPRINGER conference in Marri laxman Reddy institute of Technology, Hyderabad, Feb 2018.

[5]. Tseng, C.-J., Lu, C.-J., Chang, C.-C., Chen, G.-D., & Cheewakriangkrai, C. (2017) **Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence**. Artificial Intelligence in Medicine, 78, 47–54. https://doi.org/10.1016/j.artmed.2017.06.003

[6] Moataz M. Abdelwahab, Shimaa A. Abdelrahman,"**Four Layers Image Representation for Prediction of Lung Cancer Genetic Mutations Based on 2DPCA**", IEEE 2017, pg.no. 599-600.

[7]. Patel, S., Patel, H., (2016). **Survey of data mining techniques used in healthcare domain. International Journal of Information Sciences and Techniques** 6, 53–60.

[8] Ms. Rashmi G D, Mrs. A Lekha, Dr. Neelam Bawane, "**Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset**", IEEE Access, 2015, pg. no. 108-109.

[9]. Yang, H., & Chen, Y.-P. P. (2015) **Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information. Expert Systems with Applications**, 42(15-16), 6168–6176. https://doi.org/10.1016/j.eswa.2015.03.019

[10] G.N. Satapathi, Dr.P.Srihari, Ch.Aruna Jyothi, S. Lavanya, "**PREDICTION OF CANCER CELL USING DSP TECHNIQUES**", IEEE Access, 2013, ISBN: 978-1-4673-1622-4, pg. no. 149-151.

[11] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Information (1973-2009**), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012**, based on the November 2011 submission.

[12] Fathy, Sherif Kassem. "**A predication survival model for colorectal cancer.**" In Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and

applications, pp. 36-42. World Scientific and Engineering Academy and Society (WSEAS), 2011.

[13] Jing Wen, Jianhua Fu, Wei Zhang, and Ming Guo, "**Genetic and epigenetic changes in lung carcinoma and their clinical implications**," Modern Pathology Journal, 18 March 2011 24, pp. 932–943.

[14] S. Gupta, D. Kumar and A. Sharma, "**Data mining classification techniques applied for breast cancer diagnosis and prognosis**," Indian Journal of Computer Science and Engineering (IJCSE), Vol 2, 2011.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, **The weka data mining software: an update,** SIGKDD Explorations 11(1) (2009), 10–18.

[16] D. Chen, K. Xing, D. Henson, L. Sheng, A. Schwartz and X. Cheng, **Developing prognostic systems of cancer patients by ensemble clustering**, Journal of Biomedicine and Biotechnology 2009 (2009), 632786.
https://doi.org/10.1155/2009/632786

[17] A. Endo, T. Shibata and H. Tanaka, **Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences** 13(2) (2008), 11–16.

[18] Junkui Chen and Junzhong Gu, "**Data mining based on Colon Cancer Gene Expression Profiles,**" International Conference on Computational and Information Sciences, published by IEEE, 2007.

[19] I. Witten and E. Frank, **Data mining: Practical Machine Learning Tools and Techniques**, Morgan Kaufmann, San Francisco, CA, 2005.

[20] Abdulmajeed Alsufyani et. al., "**Detection of single-trial EEG of the neural correlates of familiar faces recognition using machine-learning algorithms** ", International Journal of Advanced Trends in Computer Science and Engineering, 8(6), November - December 2019, 2855- 2860
https://doi.org/10.30534/ijatcse/2019/28862019

[21] D. Delen, G. Walker, and A. Kadam, "**Predicting breast cancer survivability: a comparison of three data mining methods,**" Artificial intelligence in medicine, vol. 34, no. 2, pp. 113-127, 2005.

[22] M.Tawarish et al.," **An enabling technique analysis in Data Mining for Stock Market trend by Approaching Genetic Algorithm** ", International Journal of Advanced Trends in Computer Science and Engineering, 8(1), January – February 2019, 27 – 33
https://doi.org/10.30534/ijatcse/2019/06812019