ISSN No:-2456-2165

Cancer Prediction and Prognosis Using Machine Learning Techniques

Vishnu Kumar Pandey, Pankaj Pandey
Oriental Institute of Science and Technology
Bhopal, India

Abstract:- Accurate diagnosis and prediction is very important for appropriate disease treatment. Cancer is a leading cause of death worldwide, almost a million people around the globe die due to cancer every year. Cancer mortality can be reduced if it is diagnosed and treated at an early stage to save lives of cancer patients avoiding delays in care. This can be achieved with the help of machine learning. Machine learning techniques like Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), Random Forest (RMs) and Decision Trees (DTs) is broadly being used in cancer research to develop predictive models for effective and accurate prediction of cancer. We presented a review of recent ML approaches used in the modeling of cancer progression and prediction.

***Keywords**:- Machine Learning (ML), Artificial Neural Networks (ANN), Bayesian Networks (BN), Support Vector Machines (SVM), Decision Trees (DT).*

I. INTRODUCTION

Machine learning is a part of artificial intelligence which is used in different types of statistical, probabilistic and optimization techniques which helps in early diagnosis and detection of cancer. There are many machine learning algorithms like artificial neural networks (ANN) and decision trees (DT) have been used in cancer detection and diagnosis for nearly 20 years (Simes 1985; Maclin et al. 1991; Ciccheti 1992). The most common causes of cancer deaths are lung cancers, colorectal, breast, prostate etc. There are more than 100 different types of cancer and tumor, and each is categories by the type of cell that is initially affected. Breast Cancer (BC) is currently the most frequently diagnosed cancer in women [1-3]. More than 1,675,000 women are diagnosed with this disease every year and more than 500,000 die of it according to the most recent worldwide cancer data [4]. About 11,000 new cases of invasive cervical cancer are diagnosed every year in the U.S. In 2017, an estimated 15,270 children and adolescents ages 0 to 19 were diagnosed with cancer and 1,790 died of the cancers disease [5]. The overall estimate of 1,735,350 cases for 2018 equals more than 4,700 new cancer diagnoses every day. The most common cancers to be diagnosed in men are prostate, lung, and colorectal cancers, which account for 42% of all cases. The most common cancers to be diagnosed in women are breast, lung, and colorectal cancers, which combined represent one-half of all cases while breast cancer alone accounts for 30% of all new cancer diagnoses in women. In recent studies, the researchers have proven that machine learning methods could predict more accurate diagnosis or prognosis as compared to traditional statistical methods. Various machine learning techniques could be used to detect different arrangements in datasets and accordingly predict whether the cancer is benign or malignant. In this review we work on all recent uses ML techniques like Random Forest (RF), Support Vector Machine (SVM), and Bayesian Networks (BN) has been applied for the prediction of different types of cancers [6].

II. MACHINE LEARNING

The ML techniques can be divided into two main categories, supervised and unsupervised learning. In supervised learning, a set of data instances are used to train the machine and are labeled to give the correct result. However, in unsupervised learning, there are no pre-determined data sets and no notion of the expected outcome, which means that the goal is harder to achieve.

III. SURVEY OF CANCER PREDICTION

As per our survey, here we present the research publications that proposed the use of ML techniques for cancer prediction. A research has been done on the recurrence prediction of oral squamous cell carcinoma (OSCC) is proposed and they used classification algorithms for oral cancer reoccurrence [11]. They utilized different sources of data like clinical, imaging and genomic data to predict a possible relapse of OSCC and following recurrence. In this study total 86 patients were considered out of which 13 have been identified with a relapse while the remaining was disease free.

After going through all research paper, their titles and abstracts we selected only those publications in which the study of one of the three foci of cancer prediction is done and included it in their titles and in most of these studies different types of training data which includes genomic, clinical, histological, imaging, demographic, epidemiological data or combination of these are used. All the paper was excluded that focus on development of cancer prediction with the help of statistical methods like chi-square for those that use methods for cancers classification or predictive factors identification.

Machine learning and its applications are found a rapid increase in all recent papers that have been published in previous ten years [6]. Although it is impossible to achieve a complete coverage of the all literature but we tried to select a significant number of relevant paper and are presented in this review.

Machine learning helps the cancers risk assessment prediction [12-15]. We focused especially on the study published in last 5 years and their advances as compared to older publications. A machine learning based model has been developed for the assessment of women survival that is diagnosed with breast cancer [16]. A key point to several studies was that to find out the most optimal machine learning techniques from different techniques [17]. It is important to understand that in order to obtain accurate results for their prediction; the authors should select large and independent features that could result in better validation which enable extraction of more accurate and reliable predictions while it would help to minimize any bias and improve the accuracy [18].

As the authors noted that the exclusion of large number of patients due to the lack of clinical data in the research registry influenced the performance of their models. A fact that when authors used only their clinical knowledge to select 14 out of 193 variables may have resulted in significant bias and thus giving no robust results. Among the initial list of publications from our literature survey, we found a growing trend the last years regarding the prediction of cancer disease by means of SSL learning algorithm. If the quality of research studies continues to improve then the use of machine learning algorithms will become common in many clinical and hospital settings to prevent the delay of treatment [19].

Publications Relevant To ML Techniques Used For Cancer Prediction And Prognosis.

Publication	Method	Cancer type	No of patients	Type of data	Accuracy	Validation method	Important features
Exarchos K et al. [11]	BN	Oral cancer	86	Clinical, imaging tissue genomic, blood genomic	100%	10-fold cross validation	Smoker, p53 stain, extra-tumor spreading, TCAM, SOD2
Waddell M et al. [20]	SVM	Multiple myeloma	80	SNPs	71%	Leave-one-out cross validation	snp739514, snp521522, snp994532
Listgarten J et al. [21]	SVM	Breast cancer	174	SNPs	69%	20-fold cross validation	snpCY11B2 (+) 4536 T/C snpCYP1B1 (+) 4328 C/G
Stajadinovic et al. [22]	BN	Colon carcinomatosis	53	Clinical, pathologic	AUC = 0.71	Cross-validation	Primary tumor histology, nodal staging, extent of peritoneal cancer
Ayer T et al. [12]	ANN	Breast cancer	62,219	Mammographic, demographic	AUC = 0.965	10-fold cross validation	Age, mammography findings

Kim W et al. [18]	SVM	Breast cancer	679	Clinical, pathologic, epidemiologic	89%	Hold-out	Local invasion of tumor
Park C et al. [19]	Graph-based SSL algorithm	Colon cancer, breast cancer	437 374	Gene expression, PPIs	76.7% 80.7%	10-fold cross validation	BRCA1, CCND1, STAT1, CCNB1
Tseng C-J et al. [23]	SVM	Cervical cancer	168	Clinical, pathologic	68%	Hold-out	pathologic_S, pathologic_T, cell type RT target summary
Eshlaghy A et al. [17]	SVM	Breast cancer	547	Clinical, population	95%	10-fold cross validation	Age at diagnosis, age at menarche
Chen Y-C et al. [24]	ANN	Lung cancer	440	Clinical, gene expression	83.50 %	Cross validation	Sex, age, T_stage, N_stage LCK and ERBB2 genes
Park K et al. [16]	Graph-based SSL algorithm	Breast cancer	162, 500	SEER	71%	5-fold cross validation	Tumor size, age at diagnosis, number of nodes
Chang S-W et al. [25]	SVM	Oral cancer	31	Clinical, genomic	75%	Cross validation	Drink, invasion, p63 gene
Xu X et al. [26]	SVM	Breast cancer	295	Genomic	97%	Leave-one-out cross validation	50-gene signature
Gevaert O et al. [27]	BN	Breast cancer	97	Clinical, microarray	AUC = 0.851	Hold-Out	Age, angioinvasion, grade MMP9, HRASLA and RAB27B genes
Rosado P et al. [28]	SVM	Oral cancer	69	Clinical, molecular	98%	Cross validation	TNM_stage, number of recurrences
Delen D et al. [29]	DT	Breast cancer	2,00,000	SEER	93%	Cross validation	Age at diagnosis, tumor size, number of nodes, histology
Kim J et al. [30]	SSL Co-training algorithm	Breast cancer	1,62,500	SEER	76%	5-fold cross validation	Age at diagnosis, tumor size, number of nodes, extension of tumor

Table 1:- A list of published research

IV. CONCLUSION

In this review, we review the various machine learning techniques for different type of cancer prediction and prognosis (Breast Cancer, Lung Cancer, etc.). Every technique has its own accuracy and the key point in using different machine learning techniques was to find out the most optimal machine learning techniques. A comparison of various publications related to machine learning classification algorithms that aim to more accurate outcomes. We believe in that if the qualities of research studies improve, it is likely that the use of machine learning models will become much more commonplace for hospital & clinical setting to prevent the delay of treatment.

REFERENCES

[1] Youlden DR, Cramb SM, Yip CH, Baade PD. Incidence and mortality of female breast cancer in the Asia-Pacific region. Cancer biology & medicine 2014;11:101-15.
[2] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: a cancer journal for clinicians 2015;65:87-108.
[3] Cancer biology & medicine 2014.
[4] Van Grembergen O, Bizet M, de Bony EJ, Calonne E, Putmans P, Brohee S, et al. Portraying breast cancers with long noncoding RNAs. Science advances 2016;2:e1600220.
[5] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer informatics 2007;2:59-77.
[6] Nastac PJ, M. Collan, Y. Collan, T. Kuopio, B. Back. Breast cancer prediction using a neural network model. Automation Congress, 2004 Proceedings World 20 June 2005.
[7] Zafiropoulos E. MI, Anagnostopoulos I. A Support Vector Machine Approach to Breast Cancer Diagnosis and Prognosis. IFIP International Federation for Information Processing 2006.
[8] C. B. A tutorial on support vector machines for pattern recognition.
[9] B S. Statistical learning and kernel methods
[10] Exarchos KP, Goletsis Y, Fotiadis DI. Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society 2012;16:1127-34.
[11] Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Jr., Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer 2010;116:3310-21.
[12] Bochare A. GA, Yesha Y., Joshi A., Yesha Y.* Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer Int J Medical Engineering and Informatics, 2014;Vol. 6, No. 2,.
[13] Gilmore S, Hofmann-Wellenhof R, Soyer HP. A support vector machine for decision support in melanoma recognition. Experimental dermatology 2010;19:830-5.
[14] [15] Parthaláin M N. ZR. Machine learning techniques and mammographic risk assessment. International Workshop on Digital Mammography 2010.
[15] Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. Eng Appl Artif Intel 2013;26:2194-205.
[16] Ahmad LG* EA, Poorebrahimi A, Ebrahimi M and Razavi AR. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. Journal of Health & Medical Informatics April 24, 2013.
[17] Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine. J Breast Cancer 2012;15:230-8.
[18] Park C, Ahn J, Kim H, Park S. Integrative Gene Network Construction to Analyze Cancer Recurrence Using Semi-Supervised Learning. PloS one 2014;9.
[19] el. WM. Predicting Cancer Susceptibility from Single-Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma. BIOKDD 2005.
[20] Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, et al. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clinical cancer research : an official journal of the American Association for Cancer Research 2004;10:2725-37.
[21] Stojadinovic A, Nissan A, Eberhardt J, Chua TC, Pelz JO, Esquivel J. Development of a Bayesian Belief Network Model for personalized prognostic risk assessment in colon carcinomatosis. The American surgeon 2011;77:221-30.
[22] Tseng C-J. e. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Computing and Applications;24:6.
[23] Chen YC, Ke WC, Chiu HW. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med 2014;48:1-7.
[24] Chang SW, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC bioinformatics 2013;14.
[25] Li XXYZLZMWA. A gene signature for breast cancer prognosis using support vector machine. 5th International Conference on BioMedical Engineering and Informatics 2012.
[26] Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006;22:E184-E90.
[27] Rosado P, Lequerica-Fernandez P, Villallain L, Pena I, Sanchez-Lasheras F, de Vicente JC. Survival model in oral squamous cell carcinoma based on

clinicopathological parameters, molecular markers and support vector machines. Expert Syst Appl 2013;40:4770-6.
[28] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113-27.
[29] Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. J Am Med Inform Assn 2013;20:613-8.
[30] Yu Z, Lu HJ, Si HZ, Liu SH, Li XC, Gao CH, et al. A Highly Efficient Gene Expression Programming (GEP) Model for Auxiliary Diagnosis of Small Cell Lung Cancer. PloS one 2015;10.
[31] Hsia TC, Chiang HC, Chiang D, Hang LW, Tsai FJ, Chen WC. Prediction of survival in surgical unresectable lung cancer by artificial neural networks including genetic polymorphisms and clinical parameters. J Clin Lab Anal 2003;17:229-34.
[32] Santos-Garcia G, Varela G, Novoa N, Jimenez MF. Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble. Artif Intell Med 2004;30:61-9.
[33] Feng FF, Wu YM, Wu YJ, Nie GJ, Ni R. The Effect of Artificial Neural Network Model Combined with Six Tumor Markers in Auxiliary Diagnosis of Lung Cancer. J Med Syst 2012;36:2973-80.
[34] Poullis M, McShane J, Shaw M, Woolley S, Shackcloth M, Page R, et al. Lung cancer staging: a physiological update. Interact Cardiov Th 2012;14:743-9.
[35] Chatzimichail E, Matthaios D, Bouros D, Karakitsos P, Romanidis K, Kakolyris S, et al. gamma-H2AX: A Novel Prognostic Marker in a Prognosis Prediction Model of Patients with Early Operable Non-Small Cell Lung Cancer. Int J Genomics 2014.
[36] Toney LK, Vesselle HJ. Neural Networks for Nodal Staging of Non-Small Cell Lung Cancer with FDG PET and CT: Importance of Combining Uptake Values and Sizes of Nodes and Primary Tumor. Radiology 2014;270:91-8.
[37] [38] al. Ne. Early Detection of Lung Cancer Using Neural Network Techniques. Int Journal of Engineering Research and Applications 2014;4.
[38] Chen J1 CJ, Ding HY, Pan QS, Hong WD, Xu G, Yu FY, Wang YM. Use of an Artificial Neural Network to Construct a Model of Predicting Deep Fungal Infection in Lung Cancer Patients. Asian Pacific journal of cancer prevention : APJCP 2015;16:5095-9.
[39] Xie NN, Hu L, Li TH. Lung Cancer Risk Prediction Method Based on Feature Selection and Artificial Neural Network. Asian Pac J Cancer P 2014;15:10539-42.