

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/320947210>

Machine Learning Approaches in Cancer Detection and Diagnosis: Mini Review

Research · November 2017

DOI: 10.13140/RG.2.2.27775.51363

CITATION

1

READS

1,046

2 authors:



Majid Murtaza Noor

Deenbandhu Chhotu Ram University of Science and Technology, Murthal

4 PUBLICATIONS 1 CITATION

[SEE PROFILE](#)



Vinay Narwal

Maharshi Dayanand University

21 PUBLICATIONS 70 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Biomedical Equipment Modelling and Strategies for Healthcare Standards and Accrediation [View project](#)

Machine Learning Approaches in Cancer Detection and Diagnosis: Mini Review

Majid Murtaza Noor¹ and Vinay Narwal^{2*}

¹Department of Biomedical Engineering, Al Inma Medical Services, Riyadh, Saudi Arabia

²Department of Biochemistry, Maharshi Dayanand University, Rohtak, India

ABSTRACT

Cancer is the second leading cause of death globally and accounted for 8.8 million deaths in 2015. It has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. For better clinical decisions, it is important to accurately distinguish between benign and malignant tumors. Conventionally, statistical methods have been used for classification of high risk and low risk cancer, despite the complex interactions of high-dimensional medical data. To overcome the drawbacks of conventional statistical methods, machine learning has emerged as a promising technique for handling high-dimensional data, with increasing application in clinical decision support. This paper highlights new research directions and discusses main challenges related to machine learning approaches in cancer detection.

Keywords: Cancer, Machine learning, Classification, Prediction and Diagnosis.

INTRODUCTION

Cancer is not a single disease, but rather many related diseases that all involve uncontrolled cellular growth and reproduction. It is leading cause of death in the developed world and second in the developing world, killing almost 8 million people a year [1]. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. For better clinical decisions, it is important to accurately distinguish between benign and malignant tumors [2]. Conventionally, statistical methods have been used for classification of high risk and low risk cancer, despite the complex interactions of high dimensional medical data [3]. To overcome the drawbacks of conventional statistical methods, more recently machine learning has been applied to cancer prognosis and prediction [4]. Machine learning is a branch

of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to “learn” from past examples and to detect hard to discern patterns from large, noisy or complex data sets [5]. This capability is particularly well suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection. This latter approach is particularly interesting as it is part of a growing trend towards personalized, predictive medicine [6]. A number of trends are used, including a growing dependence on protein biomarkers and microarray data, a strong bias towards applications in prostate and breast cancer, and a heavy reliance on “older” technologies such artificial neural networks (ANNs) instead of more recently developed or more easily interpretable machine learning methods. A number of published studies also

appear to lack an appropriate level of validation or testing [7]. Among the better designed and validated studies it is clear that machine learning methods can be used to substantially (15–25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality. At a more fundamental level, it is also evident that machine learning is also helping to improve our basic understanding of cancer development and progression [8]. In this review different types of machine learning methods being used, the types of data being integrated and the performance of these methods in cancer prediction and prognosis.

CHALLENGES IN CLASSIFICATION OF CANCER PATIENTS

The success of modern, evidence based and personalized medical research is highly dependent on the availability of a sufficient data basis in terms of quantity and quality. This often also implies topics like exchange and consolidation of data. In the area of conflict between data privacy, institutional structures and research interests, several technical, organizational and legal challenges emerge [9]. Coping with these challenges is one of the main tasks of information management in medical research. In cancer research, case study points out the marginal conditions, requirements and peculiarities of handling research data in the context of medical research [10].

Looking at research results, it becomes obvious that cancer diseases are more like disease families with a multitude of sub-types and that the anatomical classification of tumors might be misleading and a classification according to the pathological change of signaling pathways on the cellular level is more adequate. This

differentiation is very relevant because for one patient a certain treatment may be effective and absolutely relevant while it has no positive impact on tumor control for other patients with the “same” cancer and only bears side effects [11].

In order to have an evidence-based medicine with a sound statistical basis, the amount and quality of available data becomes very important. The required amount of data increases with the number of relevant factors. Looking at the current cancer research, one has a vast array of factors and information—and it is still increasing [12, 13]. Thus, it is inevitable to cope with this heterogeneity and to build large study bases by sharing and pooling medical research data in order to realize evidence-based personalized medicine. One way to achieve this goal could be the use of machine learning techniques.

CLASSIFICATION OF MACHINE LEARNING TECHNIQUES USED IN CANCER DETECTION

Recent technologies such as microarray and next-generation sequencing have paved the way for computational methods and techniques to play critical roles in this regard. Many important problems in cell biology require the dense nonlinear interactions between functional modules to be considered. The importance of computer simulation in understanding cellular processes is now widely accepted, and a variety of simulation algorithms useful for studying certain subsystems have been designed [14]. In machine learning method, data and output is run on the computer to create a program (Fig. 1).

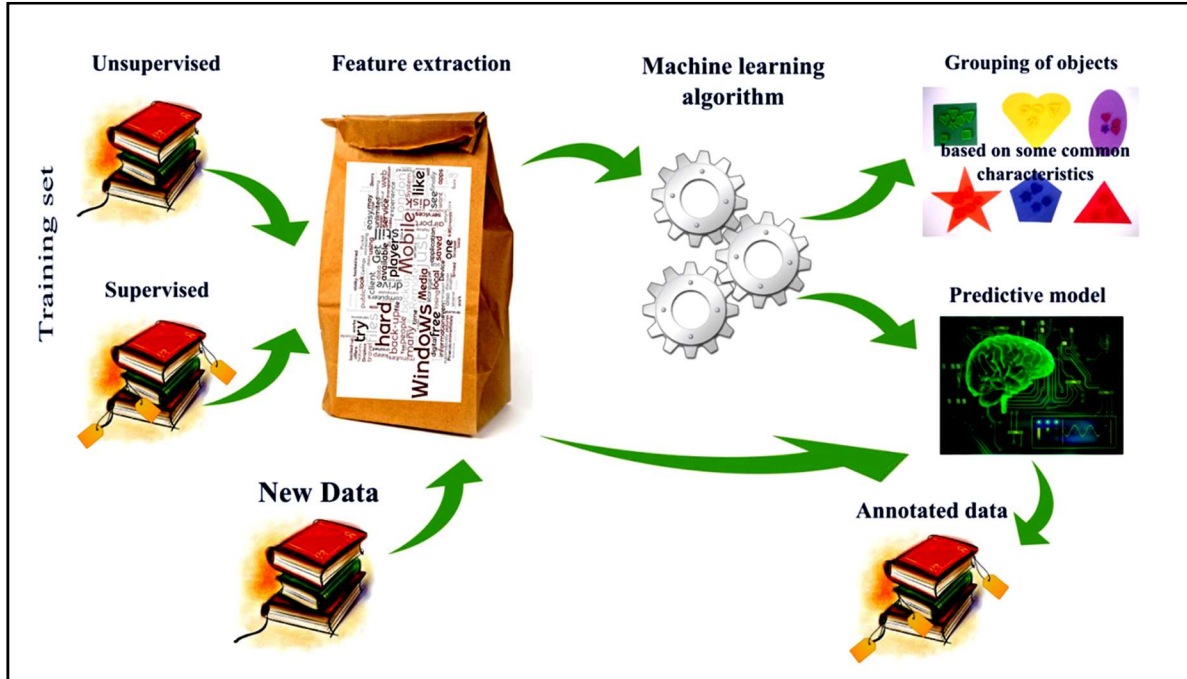


Figure: 1 Schematic representation of machine learning workflow [14].

1. Sparse compact incremental learning machine (SCILM) method

A new approach was designed to resolve cancer classification problem on microarray gene expression data, which take advantage of correntropy cost that makes it robust against diverse noises and outliers. Moreover, since SCILM uses l_1 -norm of the weights, it has sparseness, which can be applied for gene selection purposes as well. Finally, due to compact structure, the method is capable of performing classification tasks in all of the cases with only one neuron in its hidden layer. The experimental analysis in this method was performed on 26 well-known microarray data sets regarding diverse kinds of cancers and the results showed that the strategy not only achieved significantly high accuracy but also accompanied with effectivity of each gene regarding the corresponding cancer [15].

2. Knowledge base system learning method

A new knowledge-based system for classification of cancer disease was proposed using clustering, noise removal, and classification techniques. Expectation Maximization (EM) was used as a clustering method to cluster the data in similar groups. Classification and Regression Trees (CART) were used to generate the fuzzy rules to be used for the classification of cancer disease in the knowledge-based system of fuzzy rule-based reasoning method. To overcome the multi-collinearity issue, Principal Component Analysis (PCA) was incorporated in the current knowledge-based system. Experimental results on Wisconsin Diagnostic Breast Cancer and Mammographic mass datasets showed that current approach remarkably improves the prediction accuracy of breast cancer. The knowledge-based system can be used as a

3. Gauss-Newton representation based learning method

A novel Gauss-Newton representation based algorithm (GNRBA) for breast cancer classification. It uses the sparse representation with training sample selection. Until now, sparse representation has been successfully applied in pattern recognition only. This method introduces a novel Gauss-Newton based approach to find the optimal weights for the training samples for classification. In addition, it evaluates the sparsity in a computationally efficient way as compared to the conventional l_1 -norm method. The effectiveness of the GNRBA is examined on the Wisconsin Breast Cancer Database (WBCD) and the Wisconsin Diagnosis Breast Cancer (WDBC) database from the UCI Machine Learning repository. Various performance measures like classification accuracy, sensitivity, specificity, confusion matrices, a statistical test and the area

under the receiver operating characteristic (AUC) are reported

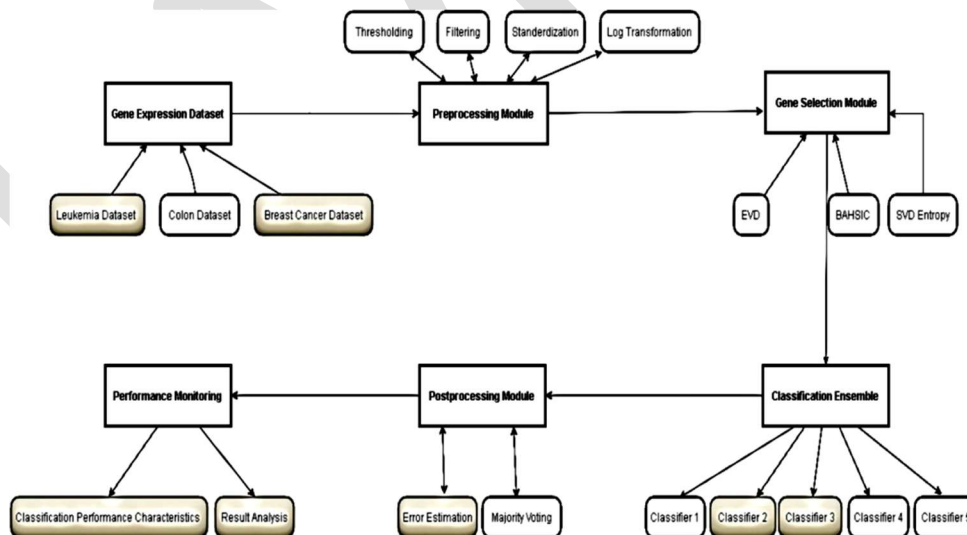


Figure: 2 Block diagram of gene expression based learning method [18].

[16].

ted to show the superiority of the proposed method as compared to classical models [17].

4. Gene expression learning method

Studying the characteristics of thousands of genes simultaneously offered a deep insight into cancer classification problem. It introduced an abundant amount of data ready to be explored. It has also been applied in a wide range of applications such as drug discovery, cancer prediction and diagnosis which is a very important issue for cancer treatment (Fig. 2). Besides, it helps in understanding the function of genes and the interaction between genes in normal and abnormal conditions [18]. That is done by monitoring the behavior of genes, gene expression data under different conditions. Ensemble classifiers increase not only the performance of the classification, but also the confidence of the results. The motivations beyond using ensemble classifiers

are that the results are less dependent on peculiarities of a single training set and because the ensemble system outperforms the performance of the best base classifier in the ensemble [19].

5. Ensemble predictive modeling framework learning method

Molecular changes often precede clinical presentation of diseases and can be useful surrogates with potential to assist in informed clinical decision making. Recent studies have demonstrated the usefulness of modeling approaches such as classification that can predict the clinical outcomes from molecular expression profiles. While useful, a majority of these approaches implicitly use all molecular markers as features in the classification process often resulting in sparse high-dimensional projection of the samples often comparable to that of the sample size. Ensemble classification approach is used for predicting good and poor-prognosis breast cancer samples from their

molecular expression profiles [20]. In contrast to traditional single and ensemble classifiers, the proposed approach uses multiple base classifiers with varying feature sets obtained from two-dimensional projection of the samples in conjunction with a majority voting strategy for predicting the class labels (Fig. 3). In contrast to earlier implementation, base classifiers in the ensembles are chosen based on maximal sensitivity and minimal redundancy by choosing only those with low average cosine distance. The resulting ensemble sets are subsequently modeled as undirected graphs. Performance of four different classification algorithms is shown to be better within the proposed ensemble framework in contrast to using them as traditional single classifier systems. [21].

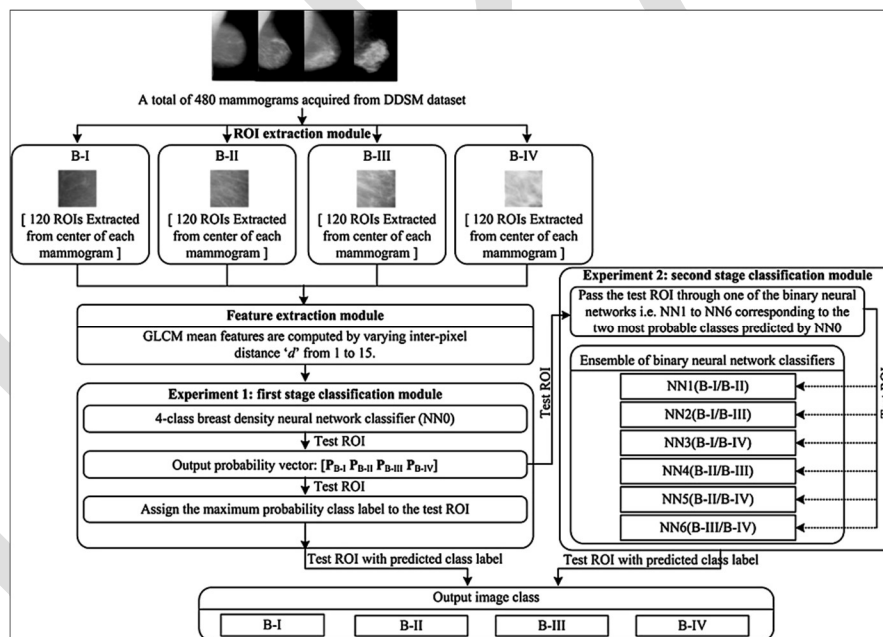


Figure: 3 Design of an efficient ensemble predictive modeling classification framework for prediction of breast density class [21].

6. Convolution neural network learning method

Deep convolutional neural networks (CNNs) were designed to obtain more discriminative description of cancer tissues. Benefiting from

the discriminative representation, metric learning layers are proposed to further improve performance of the deep structure. The best-performing model restricts the depth of back

propagation of joint training in only the metric learning layers. Relation between metric learning layers and tradition CNNs structures seems like parasitism relationship between species, where one species, the parasite, benefits at the expense of the other [22]. Therefore, the method is named as parasitic metric learning net. To confirm veracity of this method, classification experiments on breast

mass images of two widely used databases were performed (Fig. 4). Comparing performance of the current method with traditional ones, competitive results were achieved. Meanwhile, the parameter updating strategy for parasitic metric net may inspire a way of improving performance of a pre-trained CNNs model on particular medical image processing or other computer vision tasks [23].

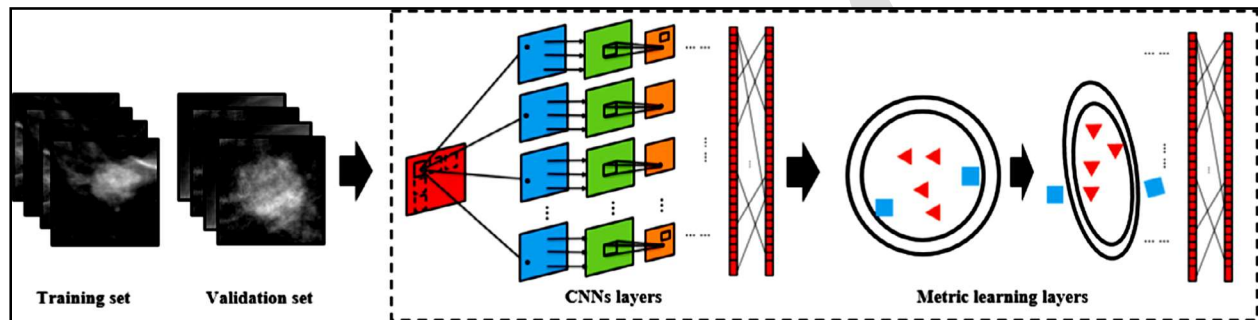


Figure: 4 Training pipeline of the proposed parasitic metric learning net [23].

CONCLUSION

In this review, we discussed the concepts of machine learning while we outlined their application in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised machine

learning methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.

REFERENCES

1. D. R. Green. (2017). Cancer and Apoptosis: Who Is Built to Last? .Cancer Cell, 31(1), 2-4.
2. H. E. Bock & U. Dold. (1967). Early Diagnosis and its Importance for Prognosis and Therapy. New Trends in the Treatment of Cancer, 1-14.
3. F. Toscani. (1996). Classification and staging of terminal cancer patients:
4. A. Bhola, & A. K. Tiwari. (2015). Machine Learning Based Approaches for Cancer Classification Using Gene Expression Data. Machine Learning and Applications: An International Journal, 2(3/4), 01-12.
5. A. H. Fielding. (1999). An introduction to machine learning methods. Machine Learning Methods for Ecological Applications, 1-35.

6. G. D. Magoulas & A. Prentza (2001). Machine Learning in Medical Applications. Machine Learning and Its Applications Lecture Notes in Computer Science, 300-307.
7. G.M. Hemphill (2016). A Review of Current Machine Learning Methods Used for Cancer Recurrence Modeling and Prediction. doi:10.2172/1329544.
8. Y.A. Malla (2017). A Machine Learning Approach for Early Prediction of Breast Cancer, International Journal Of Engineering And Computer Science. doi:10.18535/ijecs/v6i5.31.
9. J.H. Oh, R. Al-Lozi, I.E. Naqa (2009). Application of Machine Learning Techniques for Prediction of Radiation Pneumonitis in Lung Cancer Patients, 2009 International Conference on Machine Learning and Applications. doi:10.1109/icmla.2009.118.
10. J.F. Piccirillo, A.R. Feinstein (1996). Clinical symptoms and comorbidity: significance for the prognostic classification of cancer, Cancer. 77, 834–842. doi:10.1002/(sici)1097-0142(19960301)77:5<834::aid-cncr5>3.3.co;2-z.
11. A.R. Feinstein, C.K. Wells (1990). A Clinical-Severity Staging System for Patients with Lung Cancer, Medicine. 69, 1–33. doi:10.1097/00005792-199001000-00001.
12. R. Floca, (2013). Challenges of Open Data in Medical Research, Opening Science. 297–307. doi:10.1007/978-3-319-00026-8_22.
13. E.J. Topol, The big medical data miss: challenges in establishing an open medical resource, Nature Reviews Genetics. 16 (2015) 253–254. doi:10.1038/nrg3943.
14. D.W. Aha. 1991. Incremental Constructive Induction: An Instance-Based Approach, Machine Learning Proceedings 1991. 117–121. doi:10.1016/b978-1-55860-200-7.50027-1.
15. M. Nayyeri, H.S. Noghabi. 2015. Cancer Classification by Correntropy-Based Sparse Compact Incremental Learning Machine. doi:10.1101/028720.
16. M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi. 2017. A knowledge-based system for breast cancer classification using fuzzy logic method, Telematics and Informatics. 34, 133–144. doi:10.1016/j.tele.2017.01.007.
17. L. Dora, S. Agrawal, R. Panda, A. Abraham. 2017. Optimal breast cancer classification using Gauss–Newton representation based algorithm, Expert Systems with Applications. 85,134–145. doi:10.1016/j.eswa.2017.05.035.
18. S. Tarek, R.A. Elwahab, M. Shoman. 2016. Gene expression based cancer classification, Egyptian Informatics Journal. doi:10.1016/j.eij.2016.12.001.
19. X. Lu, D. Chen, 2017. Cancer classification through filtering progressive transductive support vector machine based on gene expression data. doi:10.1063/1.4992918.
20. R. Nagarajan, M. Upreti, 2017. An ensemble predictive modeling framework for breast cancer classification, Methods. doi:10.1016/j.ymeth.2017.07.011.
21. I. Kumar, B. H.s., J. Virmani, S. Thakur, 2017. A classification framework for prediction of breast density using an ensemble of neural network classifiers, Biocybernetics and Biomedical Engineering. 37, 217–228. doi:10.1016/j.bbe.2017.01.001.
22. H.H. Aghdam, E.J. Heravi. 2017. Convolutional Neural Networks, Guide to Convolutional Neural Networks, 85–130. doi:10.1007/978-3-319-57550-6_3.



I J Mutl Re App St Volume 1 Issue 1, 2017 PP. 1-8

23. Z. Jiao, X. Gao, Y. Wang, J. Li. 2017. A
parasitic metric learning net for breast
mass classification based on

mammography. Pattern Recognition.
doi:10.1016/j.patcog.2017.07.008.

UNPRAS