CANCER PRESS
www.imaqpress.com

Review Article

Machine Learning Applications in Cancer Research: Mini Review

Ghazaleh NicknamShirvan[1]

Author Information

1. Department of Network Science and Technologies, University of Tehran, Tehran,

Submitted: 29. 05.2019
Accepted: 30.05.2019
Published : 25..07.2019

Key words: Data Mining, Machine learning,Predictive Models, Interdisciplinary Researches, Cancer Researches

Abstract

Today, due to the significant growth of medical data production, the use of interdisciplinary sciences like data mining is also on rise. Data mining is a useful tool for exploring knowledge in an enormous set of medical data. One of the common data mining techniques is machine learning. This describes the ability of learning without being explicitly programmed by computers through different sets of algorithms. In the past few years, the machine learning algorithms utilization in cancer research has been the subject of growing research. This mini-review, besides defining the concepts of machine learning, reviews the application of machine learning on cancer data. The studies on this subject can be divided into four categories, including identification of high-risk people, prediction of cancer stage, prediction of cancer clinical outcomes and medical image analysis. Studies suggest the growing utilization of machine learning in medical fields and promising advancements in the future are expected.

Corresponding Author E-Mail: gh.niknam@alumni.ut.ac.ir

Introduction

Over the past decades, due to the growth of new medical technologies, the evolution of medical databases and proliferation of medical researches, an enormous amount of medical data has been produced. The investigation of such data is of great interest to researchers, but it is not possible without using statistical and computer science. Therefore, interdisciplinary studies can significantly contribute to this area. Among various subfields of computer science and statistics, one of the tools that is extensively used for examining huge sets of data is data mining. Hence, data mining has been employed by many researchers to discover the patterns or knowledge in any data groups. This process involves machine learning, artificial intelligence, statistics, and database systems. In its simplest form, data mining usually involves data preprocessing, application of data mining techniques (this study focuses on machine learning techniques) and evaluation process[1-3].

One major application of data mining is in decision support systems. Given the difficulty and complexity of decision-making process based on high-volume data, a decision support system can be of great help. A decision support system is a computerized information system used to support decision-making. This system is used in various fields of sciences and can assist physicians to make important decisions in medical science[4, 5].As mentioned earlier, today we have access to a large amount of medical data, including data from various diseases, such as cardiovascular disease, infectious diseases, cancer diseases, and many other diseases. In this context, cancer research is one of the most important issues due to the high mortality rate of cancer patients[6]. Therefore, in this article, we focuses on data mining techniques on cancer research.

In the following, first a brief overview of data preprocessing, machine learning and evaluation process is presented and then the application of these techniques in cancer researches is examined in brief.

References

1. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

2. J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

3. S. J. a. p. a. Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease," 2012.

4. D. Arnott and G. Pervan, "A critical analysis of decision support systems research revisited: the rise of design science," in Enacting Research Methods in Information Systems: Springer, 2016, pp. 43-103.

5. K. Sharma, J. J. I. J. o. A. C. Virmani, and Intelligence, A decision support system for classification of normal and medical renal disease using ultrasound images: A decision support system for medical renal diseases, International Journal of Ambient Computing and Intelligence 2017, 8(2): 52-69.

Knowledge Discovery Process

The knowledge discovery or data mining is a multi-step process. These steps are implemented in raw data to discover the desired knowledge.The overall process of knowledge discovery is presented in Figure 1. In this study, to simplify the process of knowledge discovery, it is divided into three steps of preprocessing, machine learning and evaluation process. These steps are described in the following section.

Preprocessing

It describesthe principle of examining data withhigh-quality data providinghigh-quality data analysis results. To make raw data more qualified for high qualitydata analysis, the preprocessing step should be applied. This step focuses on data modification[8]. Major tasks in data preprocessing are data cleaning, data integration, data reduction and data transformation, as shown in Figure 2[2].

Data cleaning is known as the process of detecting, correcting or removing inaccurate samples from data. Inaccurate samples include incomplete, noisy and inconsistent samples. There are severaltechniques to handle these inaccurate samples. In this study, nonetheless, only some of these techniques such as binning, regression and clustering technique are explored. As can be seen, machine learning algorithms are also utilizedat this step. Data integration techniques bring data from multiple sources into a set of instances. Then,data conflicts in this set of data are detected and resolvedby differentmethods such as correlation analysis and covariance analysis.

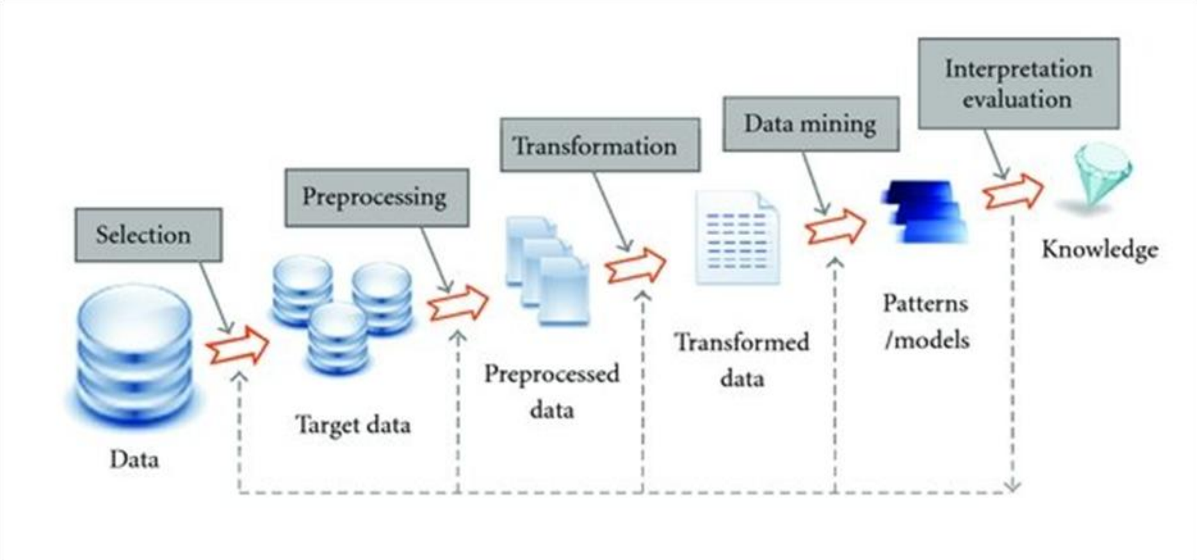


Fig1. Knowledgediscovery process[7]

References

6. L. A. Torre, R. L. Siegel, E. M. Ward, A. J. C. E. Jemal, and P. Biomarkers, Global cancer incidence and mortality rates and trends—an update. Cancer Epidemiol Biomarkers Prev. 2016; 25(1): 16-27.

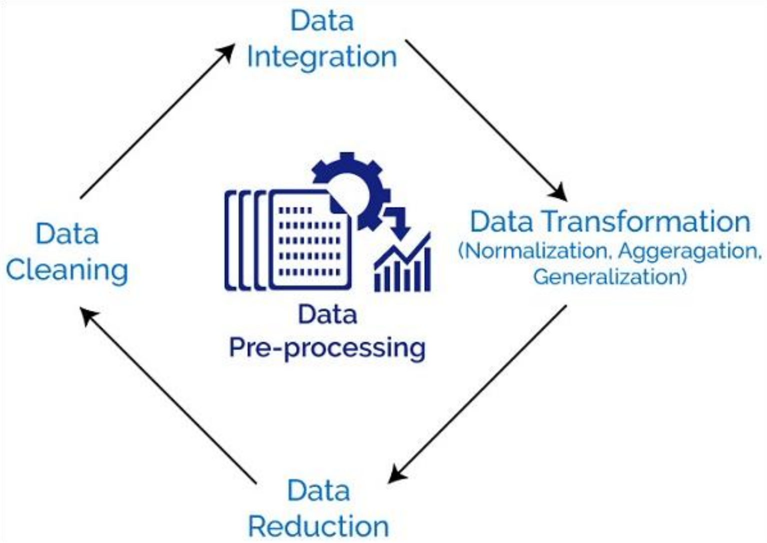
Data reduction techniques are used to reduce the amount of data that still contains critical information of the original data. Strategies ofdata reduction include dimensionality reduction, numerosity reduction and data compression. A number of techniques have been proposed for each of these strategies. For example, dimensionality reduction can use PCA and feature subset selection, among other things,andnumerosity reduction can useregression, sampling, data cube aggregation, and severalother techniques. Finally, data transformation is the process of converting data from one format or structure into another. There are variousmethod for this purpose includingnormalization and discretization.

Machine Learning

Machine learning is one of the data mining techniques thatprovides the technical basis for data mining[1]. Machine learning is a set of algorithmsthat enablescomputers to learn without being explicitly programmed.The learning process consists of two phases: i) modeling phase and ii) prediction phase. In the modeling phase, the selectedmachine learning algorithm buildsand trains a model. In the prediction phase, the model can predict new outputs[8].

It is worth noting that in machine learning algorithms, data are shown in form of tableswitheach sample being shown as a record of the table.

Fig 2. Data Preprocessing Tasks



References

7. Ö. J. A. c. i. Terzi and s. computing, Monthly rainfall estimation using data-mining process,.Applied Computational Intelligence and Soft Computing 2012; 2012: 20.

Each sample containsvariables the prediction of which is the primary goal of machine learning algorithms. This variable is referred to asclass variable, while other variables are called feature variables. Class variables are predicted based on these features. The prediction variable can be numerical or categorical and the machine learning algorithm should be selected accordingly.

Machine learning algorithms can be categorized based on their purpose. The main categories are I) Supervised Learning, II) Unsupervised Learning andIII) Semi-Supervised Learning[1, 2]. There are also other categories, such as reinforcement learning and active learning that are not addressed in this study.

Supervised Learning

In supervised learning, we label training data and they are used to build a model. In other words, sincethe actual value of the class variable of training data is known,it can be used as a guide to building the model. Some Common Algorithms of these categories are Naive Bayes, Decision Trees, Linear Regression, Support Vector Machines (SVM), Neural Networks and many other algorithms[2].

Unsupervised Learning

In this category, the model is builtwith unlabeled data, meaning that there are no output labels based on which the algorithm can model the relationships. The algorithm should be able to predict new data labels based on similarities. The main type of unsupervised learning algorithms is clustering algorithms. There are various types of clustering algorithms such as k-means, DBSCAN and chameleon clustering[2].

Semi-Supervised Learning

In semi-supervised learning, there are labeled and unlabeled data used for building a model. These types of machine learning algorithms are especially useful whendata labeling is expensive and time-consuming. In this category, with a small amount of labeled data, without spending much time and cost, the accuracy of the unsupervised learning can besignificantly enhanced [2].Among these methods, supervised and unsupervised learning are extensively used by researchers.

Evaluation Process

Evaluating the chosen machine learning algorithm is an essential part of implementing thedata mining techniques. For the evaluation process, available data is split into two subsets, training set and testing set. The modelis constructed based on thetraining set and the model performance is examined using thetestingset. The most important splitting methods areholdout, random sampling, cross-validation and bootstrap [8].There are severaltypes of evaluation metrics for supervised and unsupervised learning algorithms. In this paper, we only explore some of these metrics.With regard to theevaluation of supervised learning algorithms, Area Under a Curve (AUC), Precision and Recall and F1 Score

References

8. K. Konstantina, T. Exarchos, K. Exarchos, M. Karamouzis, D. J. C. Fotiadis,. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 2015; 13; 817.

9. Y. Wang *et al.*, Gene selection from microarray data for cancer classification—a machine learning approach. Computational Biology and Chemistry 2005; 29(1): 37-46.

are main metrics used for classifier evaluation,whileMean Absolute Error and Mean Squared Error are majormetricsused for numerical prediction. In the evaluation of unsupervised learning algorithms (i.e. clustering), evaluation methods are divided into two groups, internal methods and external methods. Internal methods such as Dunn's Index and Silhouette measure the internal correspondenceof each cluster, whileexternal methods such as Purity Index, Rand Index, and Fowlkes-Mallows measure the extent of correspondencebetween clusters.It should be noted that the aim of clustering techniques is to achieve the highest internal similarity and the lowest external similarity.

Application of Machine LearningAlgorithms in Cancer Research

The applicationof machine learning algorithms in cancer research has received growing scholarly attention in the past few years.Some of these studies such as [9-11] focus on genetic data, which are not addressedin this review.According toprevious studied, the genetic factorsnot only play an important role, but also influence theintegration of various medical data such as histological, clinical and population-based data, family history, diet, age, weight, high-risk habits in various cancer predictions [8, 12-14].The present study focuses on thesefactors.A variety of topics have been reviewed in past studies includingidentification of metastatic cancers patients [15], and prediction of possible cancer relapse [16]. In general,these studies can be assigned tofourcategories includingthe identification of people at high risk, prediction of cancer stages, prediction of cancer clinical outcomes and medical image analysis. These categories are briefly introduced in the following section.

Identification of People at High Risk

Many cancers, such as gastric cancer, have normal symptoms that are similar to other low -risk diseases. As a result, scant attention is paid to symptoms and the disease is often diagnosed at advanced stages. Unfortunately, at this stage, most treatments do not produce any significant effect, which explains the low survival rate of many cancer patients [17, 18].

It has beendemonstrated that cancer screeningforthe identification of people at high risk improves survival and reduces mortality rate[19]. For this purpose, machine learning algorithms can be applied todata collected from screening. For example, Hornbrook et al. [20]used decision trees algorithm to identify people at high risk ofcolon cancer.They used AUC to evaluate their results, with their findings yielding the highest accuracy amongother similar papers[19, 21]. Inanother example,Ayer et al. [22]applied theartificial neural network to breast cancer data.They askedsome radiologists to discriminate betweenbenign and malignant tumors .

References

10. Wang D, Li JR, Zhang YH, Chen L, Huang T, Cai YD. Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms . Genes (Basel). 2018 Mar 12;9(3).

They then drew a comparison between their own results and those reported by the radiologists, withtheir method suggestinghigh performance.In addition, they used RUC curve as the main criterion for comparing their results with thoseof previous studies. Their findings confirmed the high performance of their adopted method.

Prediction of Cancer Stage

Although the pathological stage of cancer offersthe most accurate determination of the cancerstage, surgery and its side effects are particularly harmful and undesirable in many cases. On the other hand, early prediction of cancer stage wields huge influence onthe choice of treatment decision such as urgent surgery, chemotherapy or radiation before the surgery as well asthe type of surgery. Therefore, providing a system to predict thecancerstage preoperativelyhas received increasing attention these days [23, 24].

References

11. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W . Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018 Jan-Feb;15(1):41-51.

12. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ , et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst. 2003 Mar 19;95(6):470-8.

13. Domchek SM, Eisen A, Calzone K, Stopfer J, Blackwood A, Weber BL. J Clin Oncol. 2003 Feb 15;21 (4):593-601.

14. F. Gasco, M. Valle, R. Martos, M. Zafra, R. Morales, and M. J. E. j. o. c. p. Castano. Childhood obesity and hormonal abnormalities associated with cancer risk. 2004; 13(3); 193-197.

15. Opinto G, Silvestris N, Centonze M,Graziano G, Pinto R, Fucci L, et al. Hierarchical clustering analysis identifies metastatic colorectal cancers patients with more aggressive phenotype. Oncotarget. 2017 Oct 20; 8 (50): 87782–87794 .

16. Exarchos KP, Goletsis Y, Fotiadis DI. Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans Inf Technol Biomed. 2012 Nov;16(6):1127-34.

17. Korhani Kangi A, Bahrampour A. Predicting the Survival of Gastric Cancer Patients Using Artificial and Bayesian Neural Networks. Asian Pac J Cancer Prev. 2018 Feb 26;19(2):487-490.

18. Wroblewski LE[1], Peek RM Jr, Wilson KT. Helicobacter pylori and gastric cancer: factors that modulate disease risk. Clin Microbiol Rev. 2010 Oct;23(4):713-39 .

19. Kinar Y, Akiva P, Choman E, Kariv R, Shalev V, Levin B, et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. PLoS One. 2017 Feb 9;12(2):e0171759.

They then drew a comparison between their own results and those reported by the radiologists, withtheir method suggestinghigh performance.In addition, they used RUC curve as the main criterion for comparing their results with thoseof previous studies. Their findings confirmed the high performance of their adopted method.

Prediction of Cancer Stage

Although the pathological stage of cancer offersthe most accurate determination of the cancerstage, surgery and its side effects are particularly harmful and undesirable in many cases. On the other hand, early prediction of cancer stage wields huge influence onthe choice of treatment decision such as urgent surgery, chemotherapy or radiation before the surgery as well asthe type of surgery. Therefore, providing a system to predict thecancerstage preoperativelyhas received increasing attention these days [23, 24].To this end, new studies have adoptedmachine learning methods to predict the cancer stage. The results of these studies demonstratethe effectiveness of such methods.Forexample, Pourahmad et al.[25] usedthree clustering algorithms including k-means, hierarchical and fuzzy c-means clustering methods in colorectal cancer patient to predict the cancer stage. They evaluated theresults based onexternal evaluation methods, with their findings revealing high accuracy and sensitivity of the hierarchical clustering method.Regnier-Coudert et al. [26]
usedlogistic regression, artificial neural networks, and Bayesian networks to predict

References

20. Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, et al. Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. Dig Dis Sci. 2017 Oct;62(10):2719-2727.

21. Kinar Y[1], Kalkstein N[1], Akiva P[2], Levin B[3], Half EE[4], Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. J Am Med Inform Assoc. 2016 Sep;23(5):879-90.

22. Ayer T[1], 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 Jul 15;116(14):3310-21.

23. Kehoe J, Khatri VP. Staging and prognosis of colon cancer. Surg Oncol Clin N Am. 2006 Jan;15(1):129-46.

24. Arena EA, Bilchik AJ. What is the optimal means of staging colon cancer?. Adv Surg. 2013;47:199-211.

25. Pourahmad S, Pourhashemi S, Mohammadianpanah M. Colorectal Cancer Staging Using Three Clustering Methods Based on Preoperative Clinical Finding. Asian Pac J Cancer Prev. 2016;17(2):823-7.

prostate cancer stages. They compared their results with the traditional method ofPartin tables,usingAUC as the validation criterion. The results of comparisonrevealed that BN algorithm generally excelledother methods.

Prediction of Cancer Clinical Outcomes

Early predictions of clinical outcomes can be used to improve treatment decisions [27]. The accurate prediction of these outcomes has been an interesting and challenging issue in recent years [28].There are a host ofstudies on this subject [8] most of which confirming thatthe use of machine learning techniques can improve the accuracy ofresults.Here are some of these studies that use machine learning techniques.Gründner and et al. [27] utilizedseveral machine learning algorithms to predict disease-free survival (DFS), survival, radio chemotherapy response (RCT-R), relapse, risk group stage II (SII), risk group stage III (SIII), DFS SII, relapse SII, DFS, SIII, and relapse SIII. For this purpose, they used general linear model, coxph and random forest for survival,glmnet, k-nearestneighbor, neural network, C50 decision tree, random forest and deep neural network for non-survival outcomes.Bahrani et al.[29] used C4.5 decision tree, reduced error-pruning tree, random forest, alternating decision tree, and logistic regression to predict the survival of colorectal cancer patients.Zhou and Jiang [30]used decision treesand artificial neural networks for survivability analysisof breast cancers. Delen et al. [31]performed an empiricalcomparison ofneural networks, decision trees and logistic regressionfor the prediction ofbreast cancer survivability. Chen et al. [32]used ANN for the prediction survival of non-small cell lung cancer (NSCLC) patients.

References

26. Regnier-Coudert O, McCall J, Lothian R, Lam T, McClinton S, N'dow J. Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers. Artif Intell Med. 2012 May;55(1):25-35.

27. Gründner J, Prokosch HU, Stürzl M, Croner R, Christoph J, Toddenroth D. Predicting Clinical Outcomes in Colorectal Cancer Using Machine Learning. Stud Health Technol Inform. 2018;247:101-105.

28. Zhang W, Liu Y, Sun N, Wang D, Boyd-Kirkup J, Dou X , et al.Integration of gene expression, genome wide DNA methylation, and gene networks for clinical outcome prediction in ovarian cancer. Cell Rep. 2013 Aug 15;4(3):542-53.

29. Al-Bahrani R, Agrawal A, Choudhary A. Colon cancer survival prediction using ensemble data mining on SEER data. in 2013 IEEE international conference on Big Data, 2013, pp. 9-16: IEEE.

Medical Image Analysis

Today, with the advent of new technologies in medical imaging and the numerous medical imaging modalities, accurate screening of these images poses a challenging task. Given the complexity and large quantity of these images, computer-assisted methods can be of great help to human expert in this regard[33, 34].

In light of the astonishing success of deep learning algorithms in many fields, especially image processing, this approach is used extensively for the processing of medical images. Deep learning is a machine learning technique based on artificial neural networks. Different methods of deep learning such as deep neural networks, deep belief networks, recurrent neural networks, and convolution neural networks teach computers how to do things that come naturally to humans. Deep learning is a multilayered approach that employs various layers to automatically extract features from images. It has huge potentials for the analysis of medical images[33, 35, 36]. There are an increasing number of studies in this field, of which just two cancer studies are briefly presented here. Itoh et al. [35] used convolution neural network (CNN) to investigate gastric endoscopy images. This approach capitalizes on specific features of gastric endoscopy images to detect HP infection. Early diagnosis of HP infection can help predict gastric cancer[18, 37].Itoh et al. used sensitivity, specificity and AUC as evaluation criteria. According to their results, it is possible to diagnose of HP infection by investigating gastric endoscopy images using CNN. Lundervold et al. [38]. used convolution and recurrent architectures to train a deep network. They employed this network to predict colorectal cancer outcome based on images of tumor tissue samples. Their results exhibited that prognostic information about the tissue morphology derived from deep learning techniques is even more accurate than observations

References

30. H. Zhou Z, Jiang Y. Jiang Medical diagnosis with C4. 5 rule preceded by artificial neural network ensemble. IEEE 2003; 7(1): 37-42.

31. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005 Jun;34(2):113-27.

32. Chen YC, Ke WC, Chiu HW. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med. 2014 May;48:1-7.

33. Razzak M.I, Naz S, Zaib A. Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps 2018; 26; 232-250.

34. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. ed: Elsevier, 2016.

35. Itoh T, Kawahira H, Nakashima H, Yata N. Deep learning analyzes Helicobacter pylori infection by upper gastrointestinal endoscopy image. Endosc Int Open. 2018 Feb;6(2):E139-E144.

of an experienced human observer.

Conclusion

In this study, a brief review of data mining science was under taken. To simplify the process of data mining, this process was divided into three steps of preprocessing, machine learning and evaluation process and each step was examined separately. Then, the applications of machine learning techniques in cancer research were discussed. These applications were divided into four categories of identification of people at high risk, prediction of cancer stage, prediction of cancer clinical outcomes and medical image analysis. According to these studies, recent utilization of machine learning techniques in cancer researches has been on rise, with their results revealing significant performance improvement. In addition, new developments in machine learning techniques can lead to major advances in cancer researches.

References

36. Selvikvåg Lundervold A, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik 2019; 29(2): 102-127.

37. Chen XZ, Schöttker B, Castro FA, Chen H, Zhang Y, Holleczek B , et al. Association of helicobacter pylori infection and chronic atrophic gastritis with risk of colonic, pancreatic and gastric cancer: A ten-year follow-up of the ESTHER cohort stud . Oncotarget. 2016 Mar 29;7(13):17182-93.

38. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep. 2018 Feb 21;8(1):3395.