IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 2, November 2023

Analysis of the Theory of Machine Learning in Cancer Prediction and Diagnosis of Disease

Anil Kumar Sharma[1] and Prof. (Dr.) Pushpneel Verma[2]

Research Scholar, CSE Deptt., Bhagwant University, Ajmer, Rajasthan[1]

Professor, CSE Deptt., Bhagwant University, Ajmer, Rajasthan[2]

Abstract: *Machine learning is a branch of artificial intelligence that uses a variety of statistics, probability, and optimization techniques to enable computers to "learn" from past examples and detect complex patterns from large, noisy, or complex data sets allowed. This feature is particularly suitable for clinical applications that rely on complex proteomic and genomic measurements. For this reason, machine learning is often used in cancer diagnosis and detection. Recently, machine learning has been used for cancer diagnosis and prediction. The second approach is particularly interesting because it is part of the development of personalized, predictive medicine. In writing this review, we comprehensively evaluated the different types of machine learning in use, the types of data combined, and the performance of these models in prediction and cancer diagnosis. Various assumptions have been made, including increased reliance on protein biomarkers and microarray data, bias towards prostate and breast cancer, and overreliance on "old" technology such as recently developed neural networks or simply explain machine learning. It appears that many published studies lack the necessary validation or testing. From well-designed and validated studies it is clear that machine learning techniques can be used to improve the accuracy of predicting cancer incidence, recurrence, and mortality. At a more basic level, machine learning has also been shown to improve our understanding of the onset and progression of cancer*

Keywords: AI, ML, Cancer, Onset and Progression etc

I. INTRODUCTION

The goals underlying cancer screening and diagnosis are different from cancer screening and diagnosis. Three predictive factors are involved in cancer prediction/prediction: 1) estimation of cancer incidence (i.e., risk assessment); 2) estimation of cancer incidence and 3) estimation of survival. In the first case, an attempt is made to predict the likelihood of certain types of cancer before the disease occurs. In the second case, we try to predict the outcome of the cancer after seeing the disease again. In the third part, the outcome after diagnosis (life expectancy, survival, growth, tumor drug sensitivity) is tried to be predicted. In the last two cases, the success of the prediction clearly depends in part on the success or quality of the diagnosis. However, since the diagnosis can only be made after diagnosis, the course of the disease should be considered as more than a simple diagnosis (Hagerty et al. 2005). In the future, with the help of machine learning and artificial intelligence application in cancer diagnosis, different techniques are used such as MRI, Radiography, Ultrasound, X-ray etc. Computers are also used extensively for medical training, in today's time surgeons are not only dependent on actual practice in the operation theater to acquire skills. Computers have helped a lot in the treatment, control and prevention of COVID-19, research was done on the virus of this disease through computers and later different vaccines of COVID-19 were prepared. Artificial intelligence (AI) and machine learning (ML) are gradually gaining ground in everyday life and are expected to have a major impact in digital healthcare for disease diagnosis and treatment in the near future. Technological advancements in AI and ML have paved the way towards autonomous disease diagnosis tools using large data sets to meet future challenges for early stage human disease detection, especially in cancer. ML is the subset of AI, where neural network base algorithms are developed to allow machine to learn and solve problems like human brain [1, 2]. In turn, Deep Learning (DL) is used to process data to recognize images, objects, process languages, improve drug discovery, upgrade precision medicine, improve diagnosis, and help humans make decisions. is a subset of ML to mimic the human brain's ability to It can also work without human supervision and suggest outputs [3]. DL can process data including medical images by artificial neural network

ISSN 2581-9429 IJARSCT

(ANN) to mimic human neural architecture and is composed of input, output and various hidden multi-layer networks to enhance the processing powers of machine learning. In medicine, the virtual and physical aid of technology through information management and robotics systems is the future. AI-based approaches in medicine are considered to solve complex biology puzzles, determine complex protein-protein interactions, and identify therapeutic targets. The review also discusses various trained deep-learning design models to aid in new drug discovery and robotic surgery. AI also provides medical imaging technology with extraordinary progressive potential to determine abnormal changes at the cellular level and will improve diagnostic accuracy. It also covers "AI-based precision oncology approaches" to precisely target individual cells and its role in overcoming the limitations of NGS by AI-assisted toolsets. AI-based applications in digital pathology and ethical concerns are also discussed in detail in this review to update readers about the future of medical technology.

Artificial intelligence in Diagnosis of Disease

Large technology companies, such as IBM [6] and Google, have also developed AI algorithms for healthcare. Additionally, hospitals need AI software to enable operational initiatives such as increasing cost savings, improving patient satisfaction, and meeting their staffing and workforce needs. [7] Companies are developing predictive analytics solutions that help health care managers improve business operations through increasing utilization, reducing patient boarding, reducing length of stay, and optimizing staffing levels.[8] Clinical researchers are now focusing extensively on ML algorithms, which are believed to enable computers to learn from vast pharmaceutical big data on an industrial scale, using super-computers and machine learning at low cost and in less time. Gives the ability to discover new drugs. equipment, as previously used in self-driving cars. The Exascale Compound Activity Prediction Engine (XCAPE) project, funded by Horizon 2020, a European funding program, is one of the big data analysis chemogenomic projects for chemical compound targeting biological proteins in silico models. It aims to compile comprehensive datasets of chemogenomics from authoritative databases (ChEMBL and PubChem) to predict protein interactions and gene expression for industrial scale pharmaceutical companies. ExCAPE is a scalable ML model for complex information management and its application at the industrial scale, especially in the pharmaceutical industry to predict compound biological activity and its interactions at the protein level. Nevertheless, various complex cellular limitations need to be addressed at a scalable level through algorithms and this project is expected to be further expanded by accelerating ML-based super-computers for rapid drug discovery. Recent advances in medicine for chemical synthesis include microfluidic and AI-assisted drug-designing. It has been widely proven that the trained DL-derived ML model outperformed all comparable practice strategies when applied to a database of pharmaceutical companies. What differentiates AI technology from traditional technologies in healthcare is its ability to receive information, process it, and deliver a well-defined output to the end-user. AI does this through machine learning algorithms and deep learning. These algorithms can recognize patterns in behavior and build their own logic. To reduce the margin of error, AI algorithms need to be tested repeatedly. AI algorithms behave differently from humans in two ways: (1) Algorithms are literal: if you set a goal, the algorithm cannot adjust itself and can only understand what it is explicitly told , (2) and it is not possible to explain the internal behavior of some deep learning algorithms. [1]

Machine Learning Methods

Before we begin to determine which machine learning method is best for which situation, it is important to clearly understand what machine learning is and what it is not. Machine learning is a branch of artificial intelligence that uses a variety of statistical, probabilistic, and optimization techniques to “learn” from past examples and then apply these methods. had previously been trained to classify new information, identify new patterns, or predict new patterns (Mitchell) 1997). Like statistics, machine learning is used to analyze and interpret data. But unlike statistics, machine learning can use Boolean logic (NOT, OR, NOT), conditional (IF, THEN, ELSE), conditional (result of X given Y), and negative frequency Optimize strategies to model data or classify models. . The second method is similar to the method most people use to learn and classify. Machine learning still involves a lot of statistics and probability, but it is most important because it allows decisions to be made or made that cannot be made using the traditional process (Mitchell 1997; Duda et al. 2001). For example, many statistical methods are based on multiple regression or correlation analysis. While these methods are often very powerful, they assume that the variables are independent and

ISSN
2581-9429
IJARSCT
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY

that the material can be modeled using the linear connection between these variables. Statistics often encounter problems when the relationship is nonlinear and the variables are correlated (or conditional). It's in these situations that machine learning often shines. Many biological systems are fundamentally nonlinear and their parameters depend on conditions. Many simple physical systems are linear and their parameters are essentially independent. Interestingly, almost all machine learning algorithms for cancer prediction and prediction use supervised learning. Additionally, most supervised learning algorithms exist in a special class of classes classified by probability or decision. The main types of algorithms include: 1) Artificial neural network (ANN – Rummelhart et al. 1986); 2) Decision tree (DT – Quinlan, 1986); 3) Genetic algorithm (GA – Holland 1975); 4) Linear discriminant analysis (LDA) method; 5) k-nearest neighbor algorithm estimates that more than 820 out of 1585 research papers use or refer to ANN. It was first developed by McCulloch and Pitts (1943) and later popularized by Rumelhart and others in the 1980s. (1986) states that artificial neural networks can solve many classification or pattern recognition problems. Their advantage is the ability to perform various statistics (linear, logistic and non-linear regression) and operations or assumptions (AND, OR, XOR, NOT, IF-THEN) as part of classification systems (Rodvold et al., 2001); in Chel 1997). Artificial neural networks are designed to simulate the way the brain works, where many neurons are connected to each other through many axonal connections. As in biological learning, the strength of neural connections increases or decreases through repeated training or reinforcement of educational information. Mathematically, these neural connections can be represented as a wire table or matrix (e.g. neuron 1 connects to neurons 2, 4, and 7; neuron 2 connects to neurons 1, 5, 6, and 8, etc.).

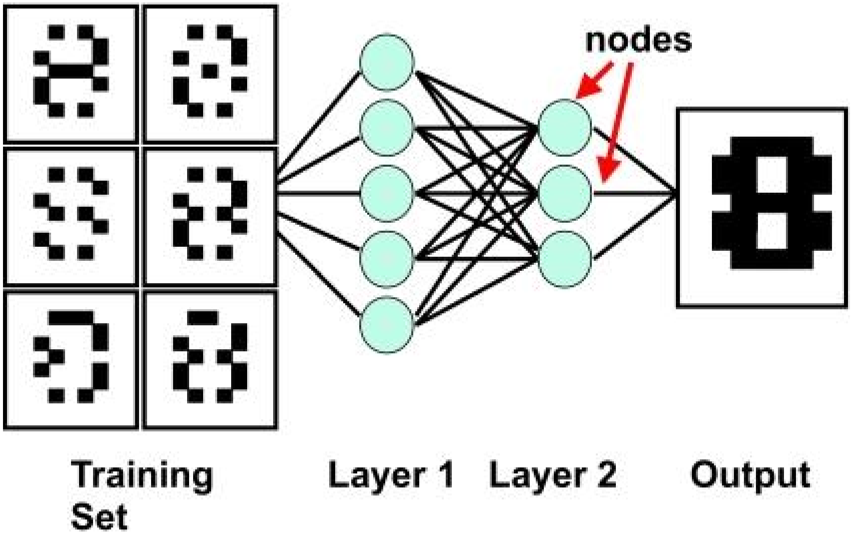

Figure 1.1: An example of training a machine learner to recognize an image written or recognized as the number "8" using the training method (negative image of the number "8"). Joseph A. Cruz and David S. Wishart (2007)

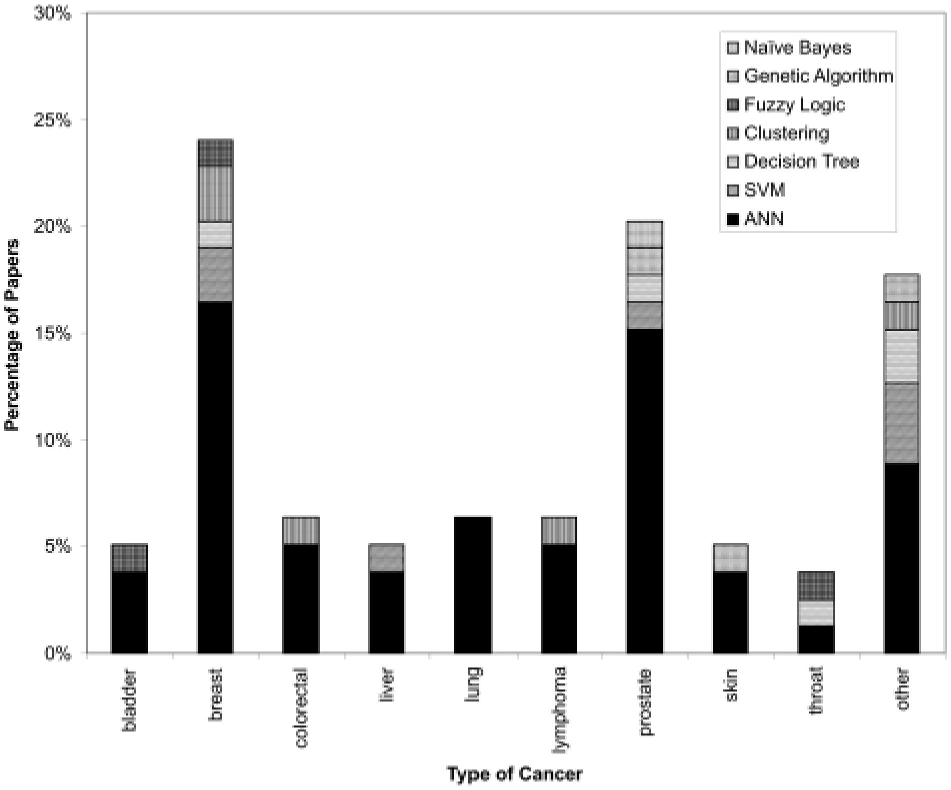

Figure 1.2: Histogram shows how well different types of machine learning are used to predict different types of cancer. Breast and prostate cancers predominate, but many cancers arising from different organs or tissues also appear to match the study's predictions. "Other" cancers include brain, cervical, esophageal, leukemia, head, neck, eye, osteosarcoma, pleural mesothelioma, breast, thyroid, and trophoblastic (uterine) malignant cancers. Joseph A. Cruz and David S. Wishart (2007)

Case Study

Cancer risk or injury prediction

Of the 79 research articles in this review, there are few articles (only 3) that use machine learning to predict cancer risk. One of the more interesting papers (Listgarten et al., 2004) developed a method to predict the development of "spontaneous" breast cancer using single nucleotide polymorphism (SNP) profiling of the enzyme that metabolizes steroids (CYP450). Familial breast cancer accounts for approximately 90% of all breast cancers (Dumitescu and Cotarla 2005). The hypothesis of this study is that certain combinations of SNPs in steroid metabolism genes increase the risk of breast cancer by causing the effects of environmental toxins or hormones in breast tissues. The authors collected SNP data (98 SNPs from 45 different cancer genes) in 63 cancer patients and 74 non-cancer patients (control group). The key to the success of this study is that the authors used different methods to reduce the sample size and learned more machine learning methods to find the best classifier. Notably, the authors quickly reduced this from the initial 98 SNPs to only 2–3 SNPs; This also seems to provide the most information. This reduces the ratio to consider 45:1 (for 3 SNPs) and 68:1 (for 2 SNPs), nowhere near 3:2 (if all 98 SNPs are used). This allowed the study to avoid falling victim to the "curse of failure" (Bellman 1961; Somorjai et al. 2016). 2003). When sample size is reduced, different types of machine learning are used, including negative Bayes models, multivariate decision tree models, and support vector machines (SVMs). SVM and Naive Bayes classifiers achieved the highest accuracy using a set of only 3 SNPs, and decision trees achieved the highest accuracy using a set of 2 SNPs. The SVM classifier performed best with 69% accuracy, while Naive Bayes and Decision Tree classifiers achieved 67% and 68% accuracy, respectively. These results are approximately 23-25% better than chance. Another difference in this research is competition and efficiency. The predictive ability of each model was analyzed in at least three ways. First, evaluate and monitor the training model with

20x cross-validation. A bootstrap resampling method was used, cross-validated 5 times, and results were averaged to minimize random events introduced into the sample distribution. Second, to reduce bias in specific selection (i.e. selection of published information on SNPs), the selection process was performed a total of 100 times in each match (5 times for each of the 20 folds). Finally, the results are compared with the non-uniform measurement test, whose measurement accuracy is more than 50%.

Cancer Survival Prediction

About half of all machine learning studies on cancer prediction focus on predicting a patient's survival rate (1-year or 5-year life expectancy). A particularly interesting paper (Fuschik et al., 2003) used a hybrid machine learning approach to predict outcomes in patients with diffuse large B-cell lymphoma (DLBCL). Specifically, clinical and genomic data (microarray) were combined to create a classification system to predict survival in patients with DLBCL. This method is slightly different from the study of Listgarten et al. (2004) used only genomic (SNP) profiles across classifier types. Fuchik et al. Predicted clinical data can supplement microarray data such that combined predictors perform better classifications based on microarray data alone or clinical data alone. The authors collected functional microarray data and clinical data from 56 DLBCL patients while collecting testing and training samples. Clinical data were obtained from the International Predictive Index (IPI), which includes risk criteria that can classify patients into less fortunate risk groups after appropriate evaluation. A simple Bayesian classifier was developed using data on the patient's IPI classification. This classification was 73.2% accurate in predicting mortality in patients with DLBCL. In contrast to Bayesian classifiers, various types of "enhanced fuzzy neural network" (EFUNN) classifiers have also been developed to process genomic data. The best EFUN classifier uses a set of 17 genes from microarray data. The accuracy of this optimized EFuNN is 78.5%.

Predicting cancer recurrence

A total of 43% of clinical studies in this review used machine learning to predict cancer recurrence or recurrence. A particularly good example is the work of De Laurentiis et al. (1999) addressed some inconsistencies reported in previous studies. These authors aimed to predict the 5-year course of breast cancer in breast cancer patients. A combination of 7 different predictors was used, including clinical data such as patient age, tumor size, and number of axillary metastases. There is also biomarker protein data, such as estrogen and progesterone receptor levels. The aim of this study is to develop a quantitative predictor that is more reliable than the lymph node metastasis (TNM) staging system. TNM is a doctor-based system based on the opinions of many doctors or experts. The authors used an ANN-based model using data from 2441 breast cancer patients (7 data points at a time), resulting in a database of more than 17,000 records. This allowed the authors to keep the model-to-feature ratio well above the recommended minimum of 5 (Somorjai et al., 2003). For optimization, all data sets are equally divided into three groups: training set (1/3), control set (1/3) and testing set (1/3) and can be used. Additionally, the authors obtained a separate group of 310 breast cancer patients from another hospital for external validation. This allowed the authors to assess the generality of their sample beyond their institution; this was a process that the two previously discussed studies did not do.

Since the aim of this study is to develop a model that predicts breast cancer better than the classical TNM staging system, it is important to compare the ANN model to TNM predictions. This was done by comparing performance using receiver operating characteristic (ROC) curves. The ANN model outperformed the TNM system (0.677) as measured by the area under the ROC curve (0.726). This work is a good example of good design and good machine learning experiments. A sufficiently large data set was obtained and data from each sample was analyzed independently to ensure validity and accuracy. Additionally, blind sets for validation were available from both the original data set and an external source to assess the generality of the machine learning models. Finally, the accuracy of the model was compared explicitly to the classical prognostic scheme, TNM staging. Perhaps a limitation of this study was that the authors tested only one type of machine learning (ANN) algorithm. Given the type and amount of data used, it is quite possible that their ANN model could outperform any other machine learning technique.

II. SUMMARY AND CONCLUSION

The size of the source data also affects the model-to-feature ratio. In general, the sample rate for each should be at least 5-10 (Somorjai et al., 2003). Small sample size/sample size is a big problem, especially for microarray studies, which often contain thousands of genes (i.e. features) but only hundreds of samples. Ohira et al. (2005) provide an example of the problems encountered when trying to overprocess microarray data. These authors developed an evidence-based database to predict outcomes in neuroblastoma patients using microarray data from 136 tumor samples. Each microarray contains 5340 genes with a sampling rate of approximately 0.025 for each trait. The feature comparison for this small sample is very sensitive to the overtraining issue. The minimum requirement for a machine learning program is a sufficiently large dataset that can be split into separate training and testing sets, or some necessary n-fold cross-validation for smaller datasets. Possible. Typically 5-fold (also using 20% of the training data as test data) or 10-fold cross-validation (also using 10% of the training data as test data) is sufficient to check courses. It is important to remember that the machine learning process is essentially a mathematical experiment. Like all experiments, it is based on theory, follows a set procedure, and needs data to implement it. Since machine learners represent real test systems, they should be treated as such. Therefore, detailed information about the process is very important. Ideally, the data used for training and testing should be described in detail and made publicly available. Information about training and testing data should be well defined, including how the data is distributed. Similarly, the algorithms used and details of their implementation should be provided or documented in a manner that allows others to verify and replicate the results. In theory, the results of a good machine learning experiment should be reproducible with other testing protocols. In particular, we analyzed some cases regarding the type of machine learning used, the type of training data combined, the type of final prediction made, the type of cancer examined and the overall performance of the method in prediction. Predisposition or occurrence of cancer. While neural networks are still powerful, it is true that alternative machine learning techniques are increasingly being used to predict at least three different outcomes in many types of cancer. It is also clear that machine learning techniques can improve the performance or prediction accuracy of most predictions, especially when compared to statistical methods or experts. While most research is generally well designed and useful, there is a need for greater experimental design and implementation, especially in terms of the quantity and quality of biological information. Improvements in experimental design and advancements in chemistry will undoubtedly increase the overall efficiency, generalizability, and reproducibility of many distributed-based systems.

REFERENCES

[1]. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230–43.

[2]. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149–53.

[3]. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Fut Healthcare J. 2019;6(2):94.

[4]. Deng L, Yu D. Deep learning: methods and applications. Found Trends Signal Process. 2014;7(3–4):197–387.

[5]. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst. 2003;95:470–8. [PubMed] [Google Scholar]

[6]. Baldus SE, Engelmann K, Hanisch FG. MUC1 and the MUCs: a family of human mucins with impact in cancer biology. Crit Rev Clin Lab Sci. 2004;41:189–231. [PubMed] [Google Scholar]

[7]. Bellman R. Adaptive Control Processes: A Guided Tour. Princeton University Press; 1961. [Google Scholar]

[8]. Bocchi L, Coppini G, Nori J, Valli G. Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. Med Eng Phys. 2004;26:303–12. [PubMed] [Google Scholar]

[9]. Bollschweiler EH, Monig SP, Hensler K, et al. Artificial neural network for prediction of lymph node metastases in gastric cancer: a phase II diagnostic study. Ann Surg Oncol. 2004;11:506–11. [PubMed] [Google Scholar]

[10]. Bottaci L, Drew PJ, Hartley JE, et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. Lancet. 1997;350:469–72. [PubMed] [Google Scholar]

ISSN 2581-9429 IJARSCT

[11]. Bryce TJ, Dewhirst MW, Floyd CE, Jr, et al. Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. Int J Radiat Oncol Biol Phys. 1998;41:239–45. [PubMed] [Google Scholar]

[12]. Burke HB, Bostwick DG, Meiers I, et al. Prostate cancer outcome: epidemiology and biostatistics. Anal Quant Cytol Histol. 2005;27:211–7. [PubMed] [Google Scholar]

[13]. Burke HB, Goodman PH, Rosen DB, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79:857–62. [PubMed] [Google Scholar]

[14]. Catto JW, Linkens DA, Abbod MF, et al. Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks. Clin Cancer Res. 2003;9:4172–7. [PubMed] [Google Scholar]

[15]. Cicchetti DV. Neural networks and diagnosis in the clinical laboratory: state of the art. Clin Chem. 1992;38:9–10. [PubMed] [Google Scholar]

[16]. Claus EB. Risk models used to counsel women for breast and ovarian cancer: a guide for clinicians. Fam Cancer. 2001;1:197–206. [PubMed] [Google Scholar]

[17]. Cochran AJ. Prediction of outcome for patients with cutaneous melanoma. Pigment Cell Res. 1997;10:162–7. [PubMed] [Google Scholar]

[18]. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273–297. [Google Scholar]

[19]. Crawford ED, Batuello JT, Snow P, et al. The use of artificial intelligence technology to predict lymph node spread in men with clinically localized prostate carcinoma. Cancer. 2000;88:2105–9. [PubMed] [Google Scholar]

[20]. Dai H, van't Veer L, Lamb J, et al. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Res. 2005;65:4059–66. [PubMed] [Google Scholar]

[21]. De Laurentiis M, De Placido S, Bianco AR, et al. A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. Clin Cancer Res. 1999;5:4133–9. [PubMed] [Google Scholar]

[22]. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34:113–27. [PubMed] [Google Scholar]

[23]. Dettling M. BagBoosting for tumor classification with gene expression data. Bioinformatics. 2004;20:3583–93. [PubMed] [Google Scholar]

[24]. Domchek SM, Eisen A, Calzone K, et al. Application of breast cancer risk prediction models in clinical practice. J Clin Oncol. 2003;21:593–601. [PubMed] [Google Scholar]

[25]. Drago GP, Setti E, Licitra L, et al. Forecasting the performance status of head and neck cancer patient treatment by an interval arithmetic pruned perceptron. IEEE Trans Biomed Eng. 2002;49:782–7. [PubMed] [Google Scholar]

[26]. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd edition. New York: Wiley; 2001. [Google Scholar]

[27]. Duffy MJ. Biochemical markers in breast cancer: which ones are clinically useful? Clin Biochem. 2001;34:347–52. [PubMed] [Google Scholar]

[28]. Duffy MJ. Predictive markers in breast and other cancers: a review. Clin Chem. 2005;51:494–503. [PubMed] [Google Scholar]

[29]. Dumitrescu RG, Cotarla I. Understanding breast cancer risk —where do we stand in 2005? J Cell Mol Med. 2005;9:208–21. [PMC free article] [PubMed] [Google Scholar]

[30]. Ehlers JP, Harbour JW. NBS1 expression as a prognostic marker in uveal melanoma. Clin Cancer Res. 2005;11:1849–53. [PubMed] [Google Scholar]

[31]. Fielding LP, Fenoglio-Preiser CM, Freedman LS. The future of prognostic factors in outcome prediction for patients with cancer. Cancer. 1992;70:2367–77. [PubMed] [Google Scholar]

ISSN
2581-9429
IJARSCT