Volume 3, Issue 3, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

LEVERAGING MACHINE LEARNING FOR PREDICTIVE ANALYTICS IN CANCER PROGNOSIS AND TREATMENT OPTIMIZATION

Wesam Taher Almagharbeh[*1], Saima Zaib[2], Misha Aslam[3], Mohammad Farrukh Tahir[4], Noor Ul Sabah[5], Syed Salah Ud Din Umer Khatab Gillani[6]

*[*1]Assistant Professor, Medical and Surgical Nursing Department, Faculty of Nursing, University of Tabuk, Tabuk Saudi Arabia 71491*

[2]MS Biomedical Engineering, NUST Isalmabad

[3]Department of Zoology, Government College University, Faisalabad

[4]Computer Engineer, Department of FCSE, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology

[5]University of Veterinary and Animal Sciences, Lahore (BS), Department of Applied Microbiology, Moscow Institute of Physics and Technology, Moscow Russia (MS), Department of Medical Biotechnology

[6]Medical Officer, DHQ Teaching Hospital Dera Ismail Khan

[*1]walmagharbeh@ut.edu.sa, [2]saima.zaib88@gmail.com, [4]farrukh.tahir@giki.edu.pk, [5]sabah.noor355@gmail.com, [6]ketogenic.123@gmail.com

[*1]https://orcid.org/0000-0002-8435-1208

DOI: https://doi.org/10.5281/zenodo.15356713

Keywords
Machine Learning, Predisposition to cancer, Forecasting models, Recurrence of cancer, Oncology prognosis.

Article History
Received on 29 March 2025
Accepted on 29 April 2025
Published on 07 May 2025

Corresponding Author: *
Wesam Taher Almagharbeh

Abstract

The medical community recognizes cancer as a multiple-disorder condition with distinct groupings. Cancer research now depends heavily on early cancer type identification and prognosis because these developments improve the subsequent therapeutic treatments for patients. Multiple research groups in biomedical and bioinformatics fields use machine learning (ML) techniques to study patient risk classification following the recognition that healthcare needs to separate subjects into high and low-risk categories. The development of neoplastic disease therapeutic approaches along with cancer development prediction utilizes these scientific methods. Computer algorithms demonstrate vital importance because they can find essential characteristics within complex datasets. Few widely applied approaches include Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), and Decision Trees (DTs which researchers use in cancer research to build predictive models for both effective and precise decision support systems. The integration of machine learning technologies in daily clinical practice requires thorough confirmation before they can be implemented due to their potential benefits in cancer development understanding. A study of present-day machine learning models that model cancer development is outlined throughout this research. This analysis contains supervised machine learning algorithms which have separate input requirements and data sets for prediction models. The increase in machine learning applications in cancer research makes it essential to analyze recently published studies that use these methods to forecast cancer chances and patient results.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

INTRODUCTION

Research into cancer has experienced continuous progress in the last few decades. Several investigative methods offered researchers the chance to recognize cancers during their initial stages before symptoms presented. Scientists have developed modern techniques to predict the results of cancer treatments during early stages. Latest medical technological developments have led to the creation of large cancer datasets available for research analysis by medical scientists. Medical staff identify defining a disease outcome with precision as a tremendous yet intriguing clinical forecasting challenge. Due to these reasons medical researchers tend to use machine learning techniques as their preferred instrument. These procedures reveal the complex relationships within complex datasets which enables accurate prediction for a particular cancer type.

Although personalized care has become necessary and machine learning has gained broader usage this article demonstrates research using these methods for cancer diagnostics and risk assessment. The study investigates prognostic and predictive indicators which either function independently from treatment choices or serve as inputs for clinical decision support for cancer patients [2]. This paper evaluates the machine learning methods employed together with their data inclusion standards and provides metrics for evaluating the proposed schemes and their strengths and weaknesses.

Multiple data types including genetic and clinical information appear in a noticeable manner throughout the proposed works. Most research studies face a common problem due to the lack of external validation testing which demonstrates how well their predictive models function. Machine learning should be implemented to boost the accuracy of cancer survival rate and risk assessment together with recurrence prediction accuracy. The implementation of machine learning approaches resulted in a 15%–20% improvement for cancer prediction accuracy in recent years as stated in [3].

Multiple studies throughout the literature introduce multiple approaches for early cancer detection and prognosis assessment as per references [4], [5], [6], [7]. The articles detail methods to study circulating miRNAs as they demonstrate promise in cancer diagnosis and identification. The detection methods exhibit restricted performance in scanning for early tumors and do not properly discriminate between benign and malignant tumor tissue. Researches about using gene expression signatures for cancer outcome predictions can be found in references [8] and [9]. The research explores both the advantages and disadvantages of microarrays when used to forecast cancer outcomes. The predictive capabilities of gene signatures for cancer patient prognostic needs do not show any clinical progress. Before clinical implementation of gene expression profiling scientists need to carry out studies testing larger and better-validated sample collections.

This paper includes research studies exclusively focused on using machine learning techniques to develop cancer prognosis models.

Machine Learning Techniques

Machine Learning operates within Artificial Intelligence as a knowledge extraction method for data samples which combines the fundamental principle of inference [10], [11], [12]. The learning system operates in two distinct stages: (i) it develops predictions about system dependencies through an examined dataset and (ii) it employs the predicted dependencies for system output forecasting. Research in biomedical science uses machine learning effectively to achieve satisfactory outcomes through the analysis of biological data sets across n-dimensional domains employing different approaches and algorithms [13]. The foremost groups within machine learning methods consist of supervised learning and unsupervised learning. Supervised learning requires input data together with labeled training data for determining specific output correlations. The process of unsupervised learning maintains no labeled examples and produces no output concept throughout its learning activities. The learning model needs to find patterns and detect clusters in incoming data streams because it lacks human direction. This situation in supervised learning represents a problem that belongs to classification types. Data classification exists as a learning methodology that produces groupings from unstructured information packs. Machine learning tasks focus primarily on regression together with clustering as fundamental practices. The learning

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

function in regression problems links input data to a real value. Using this methodology one can determine the predictive variable values for new samples. A prevalent unprovoked operation named clustering discovers groups and categories to describe information elements. Every incoming sample belongs to an identified cluster which the shared elements determine its membership.

The analysis involves studying medical breast cancer information for determining the size-related malignancy of tumors. Machine learning basically aims to predict whether a tumor is malignant (1=yes and 0=no). The process necessary to determine malignant classification of tumors is presented in Figure 1. Any errors in tumor classification because of the operation can be identified through the highlighted data.

Figure 1: Classification challenge in supervised learning. The data set contains variables X which represent tumors while their classification into benign or malignant conditions is recorded. The enclosed regions in Figure 1 depict wrongly classified cancers in the learning system.

The prevalent machine learning approach group consists of semi-supervised learning that unites supervised and unsupervised learning techniques. The learning model builds precision through the combination of labeled and unlabeled data. The approach finds regular usage when unlabeled datasets outnumber labeled datasets in a data collection.

The implementation of machine learning demands data samples that serve as its base elements. A collection of attributes defines every data sample where individual features contain multiple types of data values. Moreover, prior knowledge of the unique data type facilitates the appropriate selection of analytical tools and methodologies. The application of machine learning requires suitable preparation for enhancing raw data quality because of specific problems that arise from data-related factors. Several data quality problems exist as noise, outliers and missing or duplicate data alongside skewed or unrepresentative data. The quality of subsequent analytical work improves directly after data quality receives an enhancement. Preprocessing methods which modify raw data need to be implemented to improve its potential use for subsequent analysis. The goal of data preparation approaches is changing datasets to maximize compatibility with machine learning algorithms through various relevant methods. Dimensionality reduction stands as the most essential data preparation technique which aligns with feature selection together with feature extraction. Our method provides several key benefits for datasets which contain an extensive number of characteristics. Machine learning algorithms achieve better performance after dimension reduction takes

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

place [14]. Using dimensionality reduction allows researchers to prevent unnecessary features from the analysis while minimizing noise which leads to advanced learning models with fewer distinctive features. Feature selection methods lower data dimensions through creating new characteristics from a subset of initial characteristic groups. The embedding and filter and wrapper techniques represent the main approaches for performing feature selection according to [14]. Feature extraction establishes a new feature collection that contains all essential information in the dataset while originating from original data. Novel feature creation methods allow users to obtain the main benefits that result from dimensionality reduction.

The process of using feature selection approaches produces different predictive feature lists in specific ways. Multiple studies highlight three main issues in predictive gene matching that different research groups generate between them as well as the requirement for extensive sample numbers to achieve maximum accuracy and the lack of biological explanations for predictive patterns together with reported information leakage from published works [15, 16, 17, 18].

The main objective of machine learning methodology centers on creating models which perform classification analysis or prediction along with estimation along with related tasks. The main role in learning sequences belongs to categorization. During this function the system arranges data items into different already predefined groups. Machine learning methods enable the creation of classification models which can generate two types of errors: training-specific mistakes and generalization-related errors. Training data misclassification belongs to the first category alongside expected errors that will appear in the testing data. To achieve success as a classification model it needs to properly match the training dataset while correctly classifying every case. Test error levels advance while training error metrics decline in the situation of overfitting. According to this model condition the training mistakes which occur during model execution can reduce when the model complexity increases. A model achieves optimum results when its complexity reaches the point where the generalization error reaches its smallest possible level. The bias-variance decomposition serves as a standardized way to understand projected generalization error in learning processes. The mistake rate measurements from a particular learning method constitute its bias component. The variance of the learning process represents the secondary error source that affects all possible training data sets with fixed size and test data sets in general. The complete anticipated classification error consists of bias plus variance contributions which make up the bias–variance decomposition.

It becomes essential to implement performance assessment of your classifier which was built with machine learning approaches. Each proposed model gets performance evaluation by analyzing sensitivity, specificity, accuracy and area under the curve (AUC). A classifier demonstrates sensitivity through correct recognition of positive cases and specificity by proper identification of negative cases. The quantitative assessment of accuracy along with AUC provides the basis for evaluating the general effectiveness of a classifier. The total number of correct predictions in a model becomes measurable through specifics of accuracy. The AUC value emerges from the ROC curve to indicate how effectively the model operates as it shows sensitivity in relation to 1-specificity (Fig. 2).

Volume 3, Issue 3, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

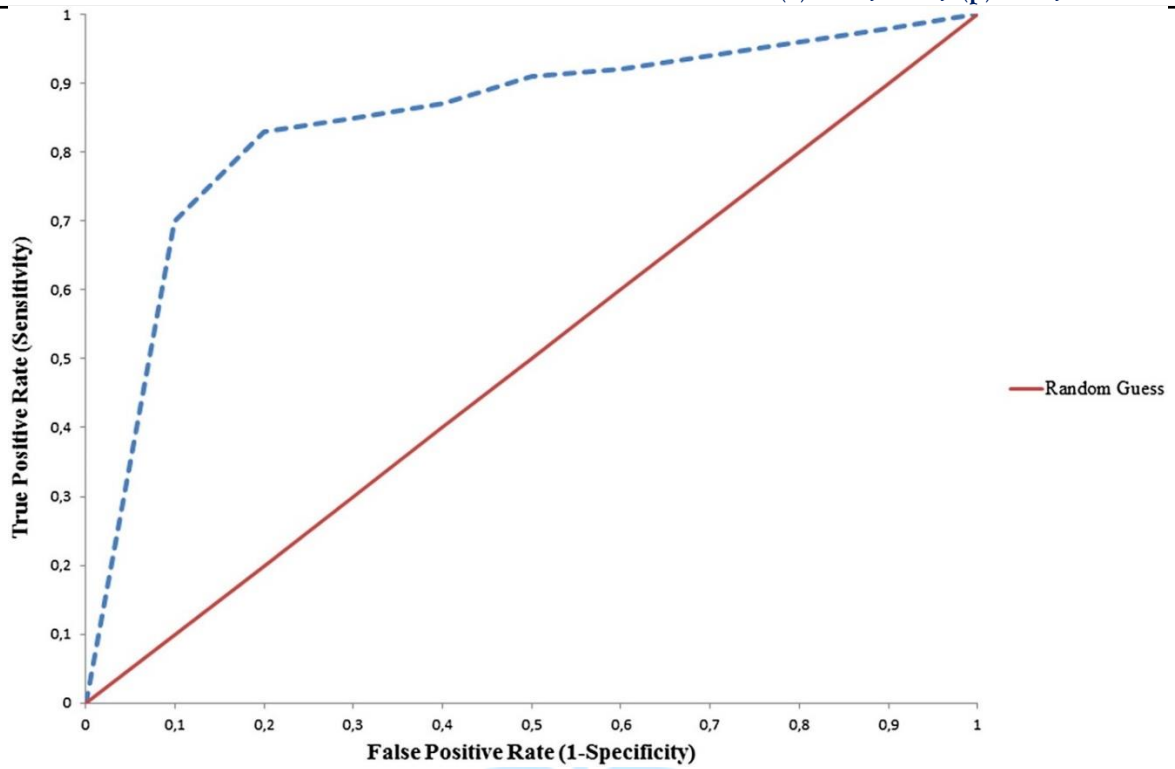


Figure 2: The illustration shows ROC curve data for two different classifiers illustrating (a) a Random Guess model with the red curve and (b) a predictive model with the blue dotted curve.

The prediction accuracy of the model originates from the testing data which allows evaluation of generalization errors. The reliable evaluation of predictive performance relies on independent training and testing sets containing sufficient data and revealing their labels so their assessment can be performed accurately. The assessment of classifier performance utilizes four established techniques which divide the original labeled dataset into parts: (i) Holdout Method, (ii) Random Sampling, (iii) Cross-Validation and (iv) Bootstrap. Data samples are divided between training and testing components during the Holdout procedure into separate sections. A classification model develops from training set data to perform effectiveness testing against test set data. The sampling technique follows similar principles as Holdout methods. The accuracy estimation process becomes stronger by implementing multiple Holdout techniques which select their training and test examples randomly each time. Cross-validation represents the third approach where all samples serve equally in training operations before being utilized only once for testing purposes. Such usage of the entire original data set leads it to be included within both the training process and testing procedure. The accuracy evaluation arises from an average calculation of every separate validation process. The bootstrap method selects training and test sets after replacement from the entire dataset so the samples get returned to the original dataset.

The selection of machine learning algorithms extends to artificial neural networks, decision trees, support vector machines and Bayesian networks after data preprocessing and learning problem definition. The research reviews machine learning procedures which scientists commonly use for cancer prediction and prognosis assessments. An examination of three main elements exists: the machine learning algorithms in use, the gathered data types and measures of evaluation applied to assess approach effectiveness in predicting cancer or analyzing illness outcomes.

The category of challenges which Artificial Neural Networks resolve includes pattern detection and classification problems. Artificial Neural Networks

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

receive input variables from which they generate outputs by uniting the received data. Many hidden layers work together to create brain-like mathematical simulations during this technique implementation. Several researchers consider artificial neural networks (ANNs) to be the standard solution for many classification problems [19] yet these systems have practical restrictions. The network design requires extensive processing time which leads to below-average results. The designated method operates within the category of "black-box" technologies. The identification of categorization mechanisms within artificial neural networks along with their failure reasons becomes nearly impossible to determine. Artificial neural networks with their linking nodes are structured as shown in Figure 3.

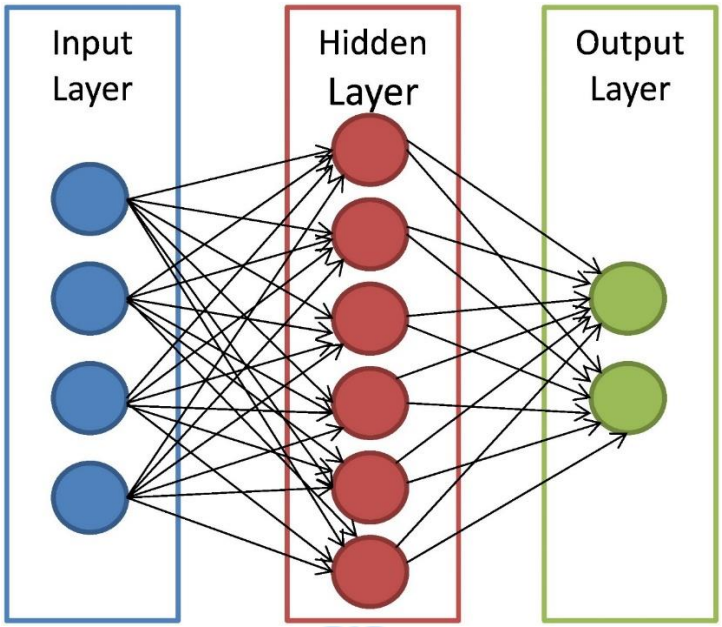


Figure 3: An example of the architecture of the artificial neural network (ANN). The arrows serve to connect nodes output with nodes input.

Decision trees arrange data within a hierarchical structure so input elements exist at nodes before leading users to decision points at leaves. Decision Trees represent one of the original primary machine learning approaches that people utilize frequently to classify data. The design of decision trees enables users to easily learn and interpret the models. The classification of a new sample can be determined by traversing through the tree. The specific design of these conclusions provides enough logical basis to make them an appealing approach. A decision tree appears in Figure 4 along with its different components and regulation elements.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

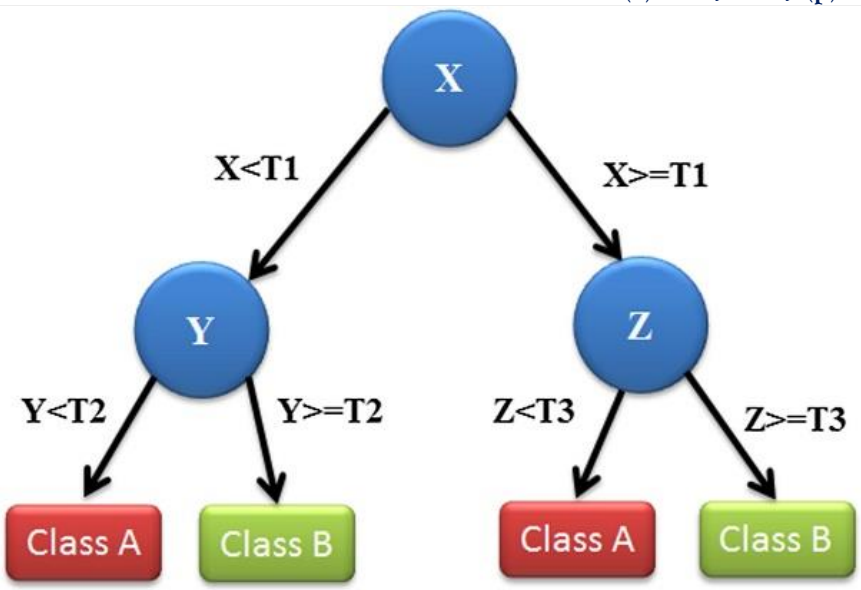
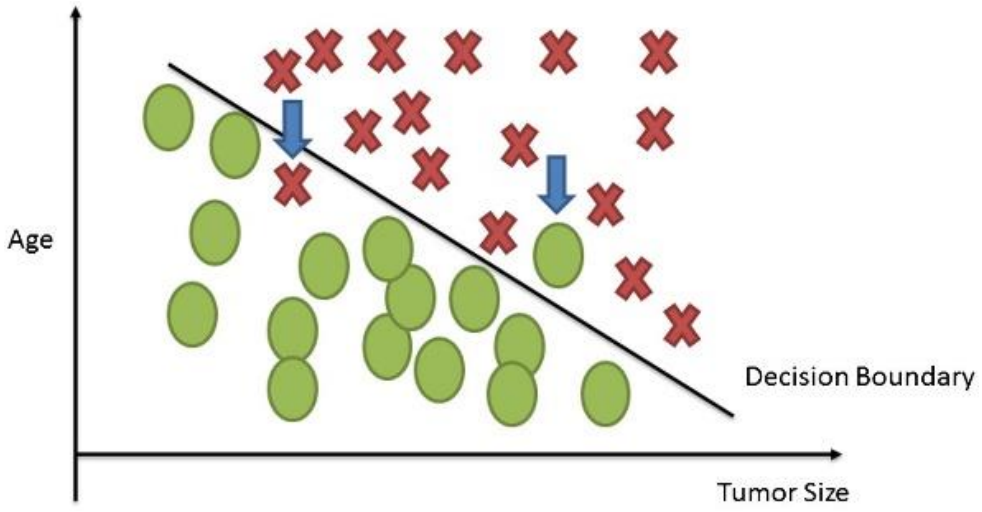


Figure 4: The following figure shows a decision tree presentation of the tree structure. The circle symbols represent the variables X, Y, Z, while choice outcomes Class A, Class B receive square symbols. The thresholds T(1–3) function as necessary criteria for proper class label classification.

The modern machine learning device Support Vector Machines (SVMs) serves as an approach to execute cancer diagnosis alongside prognosis. The conversion process for input vectors within SVMs creates elevated features before identifying a separating hyperplane for two different data clusters. A maximum distance exists between the decision hyperplane and its bordering instances. The generated classifier demonstrates strong generalization ability which allows it to serve accurately as a tool for making decisions on new data sets. SVMs enable users to generate probabilistic results according to research by [20]. The Support Vector Machine (SVM) in Figure 5 uses size and age information of tumors to identify malignant or benign tissue conditions. The created hyperplane stands as a decision frontier which serves to divide the two groupings. A decision boundary enables the detection of misclassification errors that occur during the procedure.



Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Figure 5: A simplified depiction of linear SVM categorization of the input data. The author took the illustration from the machine learning courses delivered by [21]. Tumors acquire their classification through the evaluation of patient age combined with dimensional measurements. The representation of arrows points to the tumors that the classification method misidentified.

The predictive output of Bayesian Network classifiers consists of probability values instead of concrete results. Their status indicates that Bayesian Networks serve to visualize both information and probabilistic variable relationships through directed acyclic graphs. The extensive use of Bayesian Networks exists for different classification applications and for both knowledge representation purposes and reasoning tasks.

The figure in Figure 6 features a Bayesian Network structure along with the calculated conditional probabilities among the variables.

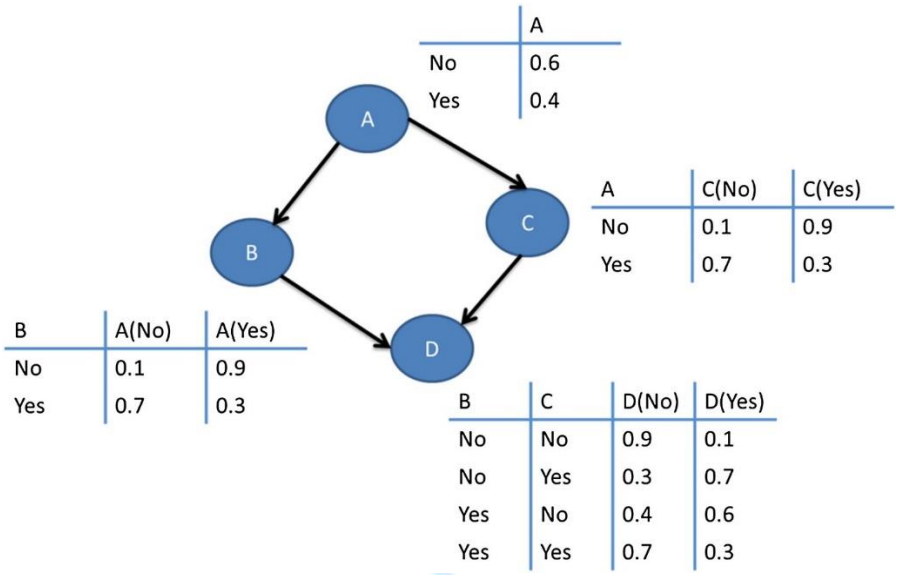


Figure 6: An example of a Bayesian Network. The collection of probabilities in each table represent random variables shown by nodes (A–D).

Machine Learning and Cancer Prediction

Since the last two decades several machine learning procedures and feature selection methods have been extensively used for medical diagnosis and prediction reporting. [3, 22, 23, 24, 25, 26, 27] Data mining techniques are used in these studies to detect cancer onset and identify major factors which lead to their incorporation in classification systems. The majority of studies employ gene expression results together with clinical characteristics alongside histological findings as integrated elements for their prognostic methodology. Machine learning algorithms serve predictions for cancer susceptibility, recurrence and survival through the studies presented in Figure 7. The scientists conducted multiple query searches which served to gather data from Scopus biomedical database. The number of publications in Fig. 3 shows the results from performing search queries combining "cancer risk assessment" AND "Machine Learning", "cancer recurrence" AND "Machine Learning", "cancer survival" AND "Machine Learning", and "cancer prediction" AND "Machine Learning". No exclusion limits were enforced on the search results yet all articles published before 2010 were taken out. Fig. 7 presents articles directly acquired from the databases yet this figure contains only minor modifications regarding publication dates.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

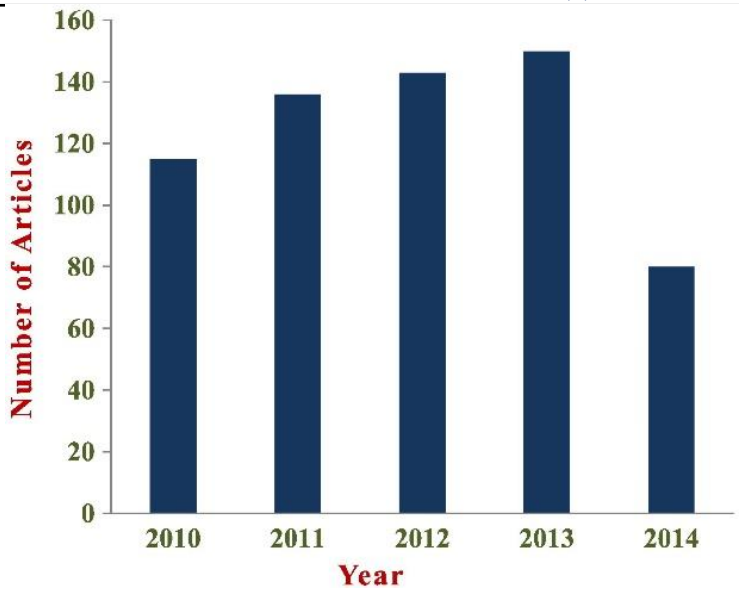


Figure 7: The number of published papers that employ machine learning methods for cancer prediction shows distribution patterns during the past five years.

A disease prognosis depends entirely on the accuracy of medical diagnosis but requires combining diagnosis data with additional information beyond it. Risk evaluation for cancer susceptibility comprises the first predictive task while cancer recurrence/local control predictions make up the second critical task with cancer survival predictions completing the set of three essential predictive areas. The initial two situations call for estimating (i) how likely someone is to develop a particular cancer form and (ii) the chance of cancer returning after achieving either full recovery or partial recovery. Survival outcome predictions that include mortality risks from cancer and overall survival times stand as the main objective in this last scenario. The field of cancer prognosis includes evaluations about life expectancy as well as survival rates and disease progression and therapy responses [3].

Successful medical diagnosis of cancer using artificial neural networks and decision trees has been practiced for nearly 30 years [22], [28], [29], [30]. Valuable research has led to 7,510 scientific articles about machine learning interactions with cancer available through PubMed. A majority of these research papers combine several machine learning algorithms with numerous data sources for tumor identification and cancer type diagnostic and prognostic purposes. The past ten years have witnessed an upward pattern of alternative supervised learning methods being utilized for cancer diagnosis and prognosis through Support Vector Machines (SVMs) and Bayesian Networks (BNs) [24], [31], [32], [33], [34], [35], [36]. Categorization algorithms have been extensively used to tackle different cancer research problems.

Standard information fitness assessment for cancer prognosis used to be based on three types of data including clinical facts population-based statistics and histologic results [23], [37]. Predicting cancer development requires consideration of personal factors such as family heritage together with age, diet, body mass, risky behaviors and carcinogenic exposure environment. [38, 39, 40]. Pertaining to only few select variables the macro-scale data enables standard statistical modeling but does not deliver enough data to reach sound conclusions. Modern technology developments in proteomics and genomics and imaging fields enable medical laboratories to obtain fresh molecular data. Multiple types of cancer-related biomarkers as well as detailed cellular patterns and gene activity demonstrate significant value for conducting cancer predictions. Modern cancer databases grew significantly through the development of High Throughput Technologies (HTTs) and now provide research access to this information. Uncertain disease predictions stand as among the most complex and intriguing issues for medical professionals during clinical practice.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Machine learning techniques have become the preferred choice for medical researchers because of their increasing popularity. This methodology identifies complex interrelationships between datasets to make accurate outlooks about the evolution of unique cancer types. Research studies have documented the use of feature selection techniques for cancer applications found in [41, 42] and [43]. The proposed computational methods identify essential aspects necessary to achieve proper illness classification.

Patients with the same cancer diagnosis currently present different subgroupings through distinct genetic patterns which lead to distinctive therapeutic options and medical results. The personalized treatment strategy depends on computational models to pick smaller affordable patient subgroups. Extensive genetic tumor type information that The Cancer Genome Atlas Research Network (TCGA) provides as a community resource can boost personalized treatments. Through high-throughput genomic technology TCGA provides researchers with advanced knowledge about cancer molecular bases.

Prediction of Cancer Susceptibility

The research was performed through Scopus and PubMed advanced searching with a time limit of five years. The research contains a study which employ machine learning methods for predicting cancer vulnerability to a particular type [55]. A genetic epidemiology investigation into bladder cancer risk is performed by the authors through Learning Classifying Systems (LCSs). The study's current case omits the genetic information work because it addresses other genetic concerns besides the main subject. The specific biological databases became the focus of our research due to the mentioned limitations. Most of the referenced literature failed to meet the fundamental requirements of either using necessary keywords from the poll or applying machine learning approaches for prediction. Among the papers we recently examined for cancer risk assessment prediction methods [19, 56, 57, 58] we found a particular study relevant to breast cancer risk estimate using artificial neural networks (ANNs) [19]. The research method differs from all other reports in this review because of its distinct choice of database type. The model adopts demographic information along with mammographic results and employs molecular and clinical along with population-based data from the selected research publications. We have decided to treat this non-selected findings study as part of our case investigation because no other results met our initial criteria. Our dismissal of the work from the general statement occurred because no other search results matched our requirements. The paper author emphasizes how researchers intensely work on building diagnostic decision tools which distinguish between breast cancer malignant versus benign symptoms. Risk stratification stands as an essential factor when developing prediction models according to their statement. Research studies analyze breast cancer patient risk using special machine learning techniques including neural networks to operate through computer models according to their findings. ANNs serve as a predictive tool for creating models which separate malignant mammographic samples from benign results. They designed their model with multiple hidden layers because they perform better generalization than networks with lesser hidden node numbers. There were 48,774 mammographic results evaluated with demographic risk factors and tumor features for analysis in this study. Professional radiologists evaluated all records of mammographic images before the corresponding interpretation data was obtained. The research team placed the dataset information into the ANN model. The validation occurred through ten-fold cross-validation evaluation methods. The authors implemented the ES method to prevent the occurrence of overfitting. During network training this approach controls errors until it spots overfitting which leads to stopping the process. The ANN achieved a calculated AUC value of 0.965 when subjected to testing and training throughout ten-fold cross-validation runs. Through their method the authors have developed a precise risk evaluation system for breast cancer patients which depends on large data sets. The ANN model training process uses primarily mammography data from tumor registry outcomes according to their team. The research employs an assessment of accuracy through the evaluation of discrimination and calibration elements. The discrimination method helps recognize malignant lesions above benign ones yet calibration serves as a tool to group

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

patients as high-risk or low-risk through risk prediction models. The authors made two analytical models: (1) They built ROC to measure discriminative power and (2) calibrated the model's performance through a calibration curve comparison to ideal breast cancer risk prediction results. The authors detected that ANNs yield identical performance when screening and diagnostic datasets are simultaneously used for input. The writers need to review the aims of preprocessing raw data for further analysis to resolve the present limitations.

Prediction of Cancer Recurrence

Our research inquiry yielded the latest relevant papers that demonstrate ML algorithms for cancer recurrence forecasting. The paper in [24] presents research about predicting oral squamous cell carcinoma (OSCC) relapses. The system needed a multiparametric Decision Support System to study how OSCC can progress after complete cancer treatment in patients. The research team applied diverse data types (clinical, imaging and genomic) to detect the potential return of OSCC which leads to recurrence development. There were 86 patients in this research who underwent evaluation with a diagnosis of recurrence found in 13 subjects and 73 showing no disease presence. The research used a particular selection framework to choose features by implementing both CFS [59] and wrapper algorithm [60]. The method prevented bias during the selection process of the most informative features from their heterogeneous reference dataset. The important variables need to be used as vectors to feed specific classifiers. The number of features consisting of clinical, imaging and genomic data equated to 65, 17 and 40 before feature selection techniques were utilized. The CFS algorithm resulted in 8 clinical, 6 imaging and 7 genomic data variables being used in each classifier after analysis. Within each classification algorithm the assessment of the smoker variable together with tumor thickness and p53 stain measurements yielded the most informative clinical variables. Using the CFS algorithm the most prominent imaging and genomic elements became the extra-tumor spreading together with the number of lymph nodes and SOD2, TCAM, OXCT2 genes.

The primary goal of this investigation involved dividing patients between those displaying disease relapse and those not displaying it before applying five classification algorithms. The proposed classification methods consist of BNs, ANNs, SVMs, DTs and RF classifiers. Each ML method received an assessment through ten-fold cross-validation as the evaluation technique for performance assessment. Accuracy alongside specificity and sensitivity values was calculated to enable comparison of the implemented classification methods. The authors analyzed ROC curves for their evaluation needs. The researchers derived their prediction outcomes from classifying data both with and without conducting feature selection and its implementation following the selection. The authors reported that a BN classifier achieved better outcomes during discrimination activities with clinical and imaging feature inputs when no feature selection was used (78.6% and 82.8% accuracy rates). Similar findings showed the best performing classifier as a BN system using CFS algorithm which achieved 91.7% accuracy. The authors created a consensus decision for detecting OSCC relapse by merging the most essential individual predictive models (BN and BN with CFS). A review of this proposal showed that it achieved superior results than other approaches used in published research. The research presented an explanatory analysis which demonstrated that mixing different data types through ML classifiers generates accurate outcomes when predicting cancer recurrence. Multiple classification techniques were used as part of their method to achieve reliable outcomes. The procedure of comparing a classifier predictor with other models leads to the discovery of the best available tool. This research has an essential limitation because it uses a small patient selection. Medical researchers analyzed only 86 disease subjects with data from their clinical examinations and image results and genomic information. The promising classification results should be evaluated against the limited number of cases when compared to the size of their data features. The research by authors from [24] published during that same year introduced BCRSVM which stands for a SVM-based model for breast cancer recurrence prediction [47]. The authors endorse how medical experts can create better care plans through risk-based patient categorization. The research investigates model development to predict breast cancer recurrence that can occur within five

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

years of surgical treatment. The authors utilized SVM alongside ANN as well as Cox-proportional hazard regression to develop their models while searching for the best model outcome. The authors obtained BCRSVM results which demonstrated better accuracy than both the other models. A piece of clinical expertise was used to choose 14 elements from the 193 variables that were available in their original dataset. The analysis consists of 733 patients drawn from 1541 selected patients led by their clinical, epidemiological and pathological variables. Kaplan–Meier analysis and Cox regression evaluation during the last phase of selection determined seven variables to be the most informative. A total of seven selected features became part of the SVM and ANN classification models and the Cox regression statistical model inputs. The authors assessed model performance by separating the data sample into two distinct subsets according to the hold-out method for training and testing. The authors evaluated the models using accuracy alongside sensitivity and specificity because this provides trustworthy results according to most studies in the literature. The authors found BCRSVM provided superior performance than both ANN and Cox regression models according to model training accuracy which achieved 84.6%, 81.4% and 72.6% respectively. BCRSVM demonstrates better performance than all other recurrence models evaluated previously. This research estimated prognostic factor significance through the use of normalized mutual information index (NMI) [61]. Among the three predictive models scientists derived the tumor's local invasion patterns as the main factor influencing breast cancer recurrence assessments. Any reviewer of this work will acknowledge specific shortcomings it contains. The authors declared the exclusion process affected model performance because they eliminated n = 808 patients from the research registry when clinical data was absent. The authors' usage of clinical knowledge to pick 14 variables out of a total of 193 potential variables appears to have created substantial bias leading to unconvincing study findings. The proposed BCRSVM model could achieve better performance through independent testing with information obtained from alternative sources by the authors. Our literature search revealed that investigations about predicting cancer via SSL learning methods experienced a growing interest during the previous years. The researchers determined the observation of current ML-based breast cancer recurrence analysis through [48] would offer valuable insight to the field. The method uses SSL to generate a graph model and joins gene expression data with gene network information for cancer recurrence prediction. The authors performed their selection of gene pairs using biological knowledge about strong biological interactions. According to the method's results the sub-gene network contains the BRCA1 gene together with CCND1 and STAT1 and CCNB1 genes. The method has three main sections that include (1) selecting gene pairs for establishing the graph model from labeled samples only then (2) constructing sample graphs based on informative genes and finally (3) applying graph regularization to identify labels of unlabeled samples. The study utilized gene expression profiles from GEO repository plus PPIs which originated from I2D database. Five datasets from the GEO platform with numbered 125, 145, 181, 249 and 111 labeled samples were obtained. The investigators classified their collected samples into three distinctive categories including recurrence and non-recurrent cases along with unattributed examples which focused on breast and colon carcinoma. A total of 194.988 human PPIs between known and experimental and predicted interactions were obtained from the I2D database by the researchers. A total of 108544 interactions remained after the process removed duplicated PPIs and those interactions without protein mapping to gene sequences. The authors used this research to demonstrate how SSL learning methodology generates networks containing critical cancer recurrence-related genes. Their technique outmatches every other current prediction method when it comes to detecting breast cancer recurrence. The proposed methodology achieved performance accuracy of 80.7% in breast cancer samples and 76.7% in colon cancer samples when compared against existing ways utilizing PPIs for informative gene detection. Cross-validation with ten folds served as the method to estimate experimental outcomes. The ML approach proves beneficial when collecting datasets due to its difference from supervised and unsupervised learning in algorithm usage while

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

offering distinct advantages regarding data collection sizes. Unlabeled data possess both low cost and simple extraction processes. Special equipment and expert personnel are needed to obtain classified data samples as opposed to their easier collection methods. The research demonstrates that SSL serves as a substitute for supervised methodologies which normally face restricted available labeled training data.

Prediction of Cancer Survival

Research in [26] develops a prediction model to forecast breast cancer survival in women by demonstrating how resilient factors represent critical influences on the model variables. The authors used SEER cancer database [64] to test three classification models including SVM, ANN and SSL. Sixteen main attributes exist within the dataset which includes 162,500 records. The research studied survival as a class variable that differentiates between patients who died from those who survived. The three most significant parameters in diagnosis include tumor size together with the number of nodes and age at diagnosis. The optimal performance evaluation showed ANN achieved 65% accuracy while SVM reached 51% but SSL delivered 71%. The prediction models received an assessment through five-fold cross-validation. Clinical professionals could implement the SSL model because of these results according to the authors. No initial handling techniques were mentioned in the article for selecting the best traits during their data collection phase. The analysts applied the full SEER database through a box-whisper-plot for parameter-performance analysis across 25 model parameter options. A model demonstrates high stability and resilience whenever its specified parameters produce compact box regions. The SSL design presented more precise performance outcomes within its specific box pattern than the other SSL model types did. Researchers evaluated NSCLC patient survival predictions by applying ANNs in a study published during the subsequent year [50]. NSCLC patients' raw gene expression data and clinical information are obtained from the NCI caArray database [65] which forms the basis of the study dataset. Scientists completed their methodology preprocessing stages before selecting both LCK and ERBB2 genes because these showed survival association and were later employed to train their ANN network. The ANN model received four clinical input variables that included age, T_stage, N_stage together with sex. The research team investigated many ANN designs to determine which model performed best at cancer survival prediction. The categorization system reached 83% success rate as its prediction accuracy measure. The study results showed that all patients receiving treatment belonged to particular groups and 50% of these patients experienced death. Survival analysis procedures based on Kaplan–Meier provided the assessment of the model outcome. Three sets of patient survival data received assessment using the index of $p < 0.00001$ which indicated that patients identified as high-risk experienced shorter median overall survival compared to individuals classified in low-risk categories. The research provided more uniform survival prediction results for NSCLC when compared to past studies concerning the same subject. The present article faces a significant shortcoming because it does not include blood clots and other death-related factors which could lead to misdiagnosis results. The paper states its methodology limited to NSCLC cancer and unsuitable for different cancer types. The current research faces a notable restriction because researchers make the assumption that prediction models will not apply consistently to various cancer types.

Discussion

The review examines recent research involving ML methods for cancer prediction and prognosis activities. A preliminary introduction to the ML framework followed by explanations of preprocessing data approaches and selection features methods and classification techniques leads to the presentation of three exemplary studies which use prevalent ML tools for cancer risk prediction and cancer recurrence assessment and cancer survival prediction. An abundant number of accurate ML studies about specific cancer outcome predictions have emerged in the last ten years. The extraction process for clinical decisions needs an assessment of potential issues that stem from experimental design choices and data

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

collection methods along with the validation methods of classified results.

The claims about effective and adequate decision making through ML classification techniques remain unproven because these techniques have failed to penetrate actual clinical practice. The development of modern omics technologies created new ways to understand different diseases but complete validation results are required to effectively implement gene expression signatures in medical settings.

Studies during the last 2 years increasingly applied semi-supervised ML techniques for developing cancer survival models. The prediction process with this algorithm combines labeled together with unlabeled data elements which research has found to produce better outcomes than traditional supervisory techniques [26]. SSL offers an excellent substitute for traditional supervised ML and unsupervised ML approaches since it operates with restricted labeled examples.

According to the reviewed studies this review discovered that a primary constraint exists due to minimal available data samples. All categorization systems in disease modeling must have training datasets which reach a required minimum magnitude for adequate performance. Having an abundance of available data enables researchers to create separate training and testing areas which provides reliable validation of their estimators. A training dataset that is smaller than the dimension of the data will lead to model misclassification and the estimators will produce biased and unstable results. Using a wider group of patients in survival prediction leads to better overall model applicability.

Neither data volume nor data quality alongside well-designed feature selection methods should be ignored in achieving effective ML for accurate cancer predictions. Advancing performance outcomes in predictive models happens when using feature selection techniques to identify the most beneficial feature subset. Other feature sets whose components include histological testing and pathological evaluations display quantifiable value measurements. When working with clinical variables that lack static properties it becomes important for ML systems to adapt their operations to different feature collections over time.

The works included nearly every validation assessment procedure to estimate their learning system performance. Researchers used recognized evaluation methods that caused the primary datasets to divide into multiple sections. The research team needs to pick deep and separate feature collections to get precise predictive outcomes according to previously mentioned guidelines. Internal together with external validation procedures were present in these research works which delivered better predictability as well as minimized bias [47].

The research success of several studies depended on using multiple ML approaches to discover their optimal solution [34]. The trend involves feeding models with various types of data in combination for prediction purposes. During the past decade researchers only utilized information from molecular activities and clinical data to forecast cancer outcomes. Rapid HTT advancement which includes genomic, proteomic along with imaging technologies led to the development of new types of input parameters. Research showed that predictions methods required integration of genomic, clinical, histological, imaging demographic, epidemiological data and proteomic data. The predictions also combined various subsets from the same data groups [24, 26, 48, 50, 53].

Significant research exists about combining different data types within the domain of breast cancer analysis [66,67]. Clinical treatment scores were integrated with signatures based on immunohistochemistry as well as expression-based signatures including PAM50 and Oncotype DX within the DREAM project [68] to show substantial advancement of treatment selection by combining various features.

The prediction outcomes for cancer patients utilize SVM and ANN classifiers as the most common applied ML algorithms. The research literature reveals that ANNs were introduced into practice almost three decades ago [30]. SVMs represent recent technology in cancer prediction/prognosis and they continue to find broad applications because of their precise results. Various parameters such as data collection types and samples sizes and operational time constraints determine which algorithm will be the most suitable.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Researchers need to examine innovative cancer modeling techniques for addressing the difficulties outlined previously. The use of improved statistical methods to analyze diverse datasets would yield precise results together with explanations regarding disease progression. More public databases need to be constructed to gather valid cancer dataset from all patients who have received the disease diagnosis. The researchers can use clinically exploited datasets to develop better models that would lead to improved model validity and clinical decision integration.

Conclusion

The paper investigates machine learning principles for cancer prognosis and prediction applications. Several scientific studies during recent years focus on creating prediction models through supervised machine learning techniques with classification algorithms for precise illness predictions. Various data integration techniques alongside multiple selection approaches and classification methods show promise for cancer inference applications according to their evaluated research findings.

REFERENCES

[1] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144: 646–74.

[2] Polley M-YC, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. J Natl Cancer Inst 2013;105:1677–83.

[3] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informat 2006;2:59.

[4] Fortunato O, Boeri M, Verri C, Conte D, Mensah M, Suatoni P, et al. Assessment of circulating microRNAs in plasma of lung cancer patients. Molecules 2014;19:3038–54.

[5] Heneghan HM, Miller N, Kerin MJ. MiRNAs as biomarkers and therapeutic targets in cancer. Curr Opin Pharmacol 2010;10:543–50.

[6] Madhavan D, Cuk K, Burwinkel B, Yang R. Cancer diagnosis and prognosis decoded by blood-based circulating microRNA signatures. Front Genet 2013;4.

[7] Zen K, Zhang CY. Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers. Med Res Rev 2012;32:326–48.

[8] Koscielny S. Why most gene expression signatures of tumors have not been useful in the clinic. Sci Transl Med 2010;2 [14 ps12-14 ps12].

[9] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92.

[10] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[11] Mitchell TM. The discipline of machine learning: Carnegie Mellon University. Carnegie Mellon University, School of Computer Science, Machine Learning Department; 2006.

[12] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2005.

[13] Niknejad A, Petrovic D. Introduction to computational intelligence techniques and areas of their applications in medicine. Med Appl Artif Intell 2013;51.

[14] Pang-Ning T, Steinbach M, Kumar V. Introduction to data mining; 2006.

[15] Drier Y, Domany E. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? PLoS One 2011;6:e17795.

[16] Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147–57.

[17] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21:171–8.

[18] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci 2006;103:5923–8.

[19] Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited. Cancer 2010;116:3310–21.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

[20] Platt JC, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification; 1999 547–53.

[21] Adams S. Is Coursera the beginning of the end for traditional higher education? Higher Education; 2012.

[22] Cicchetti D. Neural networks and diagnosis in the clinical laboratory: state of the art. Clin Chem 1992;38:9–10.

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

[24] Exarchos KP, Goletsis Y, Fotiadis DI. Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans Inf Technol Biomed 2012;16: 1127–34.

[25] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89–109.

[26] Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. Engl Appl Artif Intell 2013;26:2194–205.

[27] Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics 2007;23:30–7.

[28] Bottaci L, Drew PJ, Hartley JE, Hadfield MB, Farouk R, Lee PWR, et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. Lancet 1997;350:469–72.

[29] Maclin PS, Dempsey J, Brooks J, Rand J. Using neural networks to diagnose cancer. J Med Syst 1991;15:11–9.

[30] Simes RJ. Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. J Chronic Dis 1985;38:171–86.

[31] Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 2009;36:3240–7.

[32] Chang S-W, 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 Bioinforma 2013;14:170.

[33] Chuang L-Y, Wu K-C, Chang H-W, Yang C-H. Support vector machine-based prediction for oral cancer using four snps in DNA repair genes; 2011 16–8.

[34] Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, Ahmad LG. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform 2013;4:124.

[35] Exarchos KP, Goletsis Y, Fotiadis DI. A multiscale and multiparametric approach for modeling the progression of oral cancer. BMC Med Inform Decis Mak 2012;12:136.

[36] Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. J Am Med Inform Assoc 2013;20:613–8.

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

[38] 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;95:470–8.

[39] Domchek SM, Eisen A, Calzone K, Stopfer J, Blackwood A, Weber BL. Application of breast cancer risk prediction models in clinical practice. J Clin Oncol 2003;21: 593–601.

[40] Gascon F, Valle M, Martos R, Zafra M, Morales R, Castano MA. Childhood obesity and hormonal abnormalities associated with cancer risk. Eur J Cancer Prev 2004;13: 193–7.

[41] Ren X, Wang Y, Chen L, Zhang X-S, Jin Q. ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. Nucleic Acids Res 2013;41:e53.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

[42] Ren X, Wang Y, Zhang X-S, Jin Q. iPcc: a novel feature extraction method for accurate disease class discovery and prediction. Nucleic Acids Res 2013:gkt343.

[43] Wang Y, Wu Q-F, Chen C, Wu L-Y, Yan X-Z, Yu S-G, et al. Revealing metabolite biomarkers for acupuncture treatment by linear programming based feature selection. BMC Syst Biol 2012;6:S15.

[44] Waddell M, Page D, Shaughnessy Jr J. Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma. ACM 2005:21–8.

[45] 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. Clin Cancer Res 2004;10:2725–37.

[46] Stojadinovic A, Nissan A, Eberhardt J, Chua TC, Pelz JOW, Esquivel J. Development of a Bayesian belief network model for personalized prognostic risk assessment in colon carcinomatosis. Am Surg 2011;77:221–30.

[47] Kim W, Kim KS, Lee JE, Noh D-Y, Kim S-W, Jung YS, et al. Development of novel breast cancer recurrence prediction model using support vector machine. J Breast Cancer 2012;15:230–8.

[48] Park C, Ahn J, Kim H, Park S. Integrative gene network construction to analyze cancer recurrence using semi-supervised learning. PLoS One 2014;9:e86309.

[49] Tseng C-J, Lu C-J, Chang C-C, Chen G-D. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput & Applic 2014;24: 1311–6.

[50] Chen Y-C, Ke W-C, Chiu H-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med 2014;48:1–7.

[51] Xu X, Zhang Y, Zou L, Wang M, Li A. A gene signature for breast cancer prognosis using support vector machine. IEEE 2012:928–31.

[52] 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:e184–90.

[53] Rosado P, Lequerica-Fernández P, Villallaín L, Peña 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.

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

[55] Urbanowicz RJ, Andrew AS, Karagas MR, Moore JH. Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: a learning classifier system approach. J Am Med Inform Assoc 2013;20:603–12.

[56] Bochare A, Gangopadhyay A, Yesha Y, Joshi A, Yesha Y, Brady M, et al. Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer. Int J Med Eng Inform 2014;6:87–99.

[57] Gilmore S, Hofmann-Wellenhof R, Soyer HP. A support vector machine for decision support in melanoma recognition. Exp Dermatol 2010;19:830–5.

[58] Mac Parthaláin N, Zwiggelaar R. Machine learning techniques and mammographic risk assessment. Digital mammography. Springer; 2010. pp. 664–672.

[59] Hall MA. Feature selection for discrete and numeric class machine learning; 1999. [60] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97: 273–324.

[61] Estévez PA, Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. IEEE Trans Neural Netw 2009;20:189–201.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

[62] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles — database and tools update. Nucleic Acids Res 2007;35:D760–5.

[63] Niu Y, Otasek D, Jurisica I. Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. Bioinformatics 2010;26:111–9.

[64] Howlader N, Noone A, Krapcho M, Garshell J, Neyman N, Aletkruse S. SEER Cancer Statistics Review, 1975–2010, [Online] National Cancer Institute. Bethesda, MD: National Cancer Institute; 2013 [Online].

[65] Bian X, Klemm J, Basu A, Hadfield J, Srinivasa R, Parnell T, et al. Data submission and curation for caArray, a standard based microarray data repository system; 2009.

[66] Papadopoulos A, Fotiadis DI, Costaridou L. Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques. Comput Biol Med 2008;38:1045–55.

[67] Papadopoulos A, Fotiadis DI, Likas A. Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines. Artif Intell Med 2005;34:141–50.

[68] Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. PLoS Comput Biol 2013;9:e1003047.

[69] Cuzick J, Dowsett M, Pineda S, Wale C, Salter J, Quinn E, et al. Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. J Clin Oncol 2011;29:4273–8.

[70] Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7.

[71] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004; 351:2817–26.