Al-Salam Journal for Engineering and Technology
Journal Homepage: http://journal.alsalam.edu.iq/index.php/ajest
e-ISSN: 2790-4822 p-ISSN: 2958-0862

Predicting of cancer disease using machine learning models, a review

Oqbah Salim Atiyah[1]*

[1]Computer Science Department, Computer Sciences College, Tikrit University, Salah-Aldeen, 34001, Iraq.

*Corresponding Author: Oqbah Salim Atiyah

DOI: https://doi.org/10.55145/ajest.2025.04.01.009
Received June 2024; Accepted August 2024; Available online September 2024

ABSTRACT: Cancer is a serious disease that leads to mortality if not exposed at a precocious phase. This is what makes it worrying, as no actual treatment has been found for this disease yet, and patients with the disease cannot be saved unless it is diagnosed early in the first and second stages. But, if it is discovered in the third and fourth stages, the patient has little chance of survival. This is why using machine learning is a beneficial technology to deal with these diseases to help in their early detection. Machine learning models effectively learn from a training dataset how to estimate entire processes. Many models are offered by machine learning to forecast cancer based on standard datasets gathered by healthcare websites, social media, and some other repositories. That is why there is an urgent need to apply machine learning algorithms to this dataset for early cancer detection in affected people. The essential goal of the review is to highlight and estimate the best approaches used and make it easier for other researchers to conduct relevant studies utilizing machine learning models on cancer. The survey indicates that the models are competitive and demonstrate the accuracy of the classifiers on the datasets they were trained on.

Keywords: Cancer datasets, Machine Learning models, Patient features

1. INTRODUCTION

World Health Organization announced that cancer is a serious disease and the second largest disease killer of humans [1]. The cancer is a serious illness featuring cell growth extraordinary within the body over which has no control. It spreads at high speed and affects further portions of the body. gradually, the cancer damages the person's body when extraordinary growth of cells forms many masses of tissue in the person's body, and named with tumors. some cancerous tumors do not appear in the person's body. It may have expanded and engaged with further portions such as the circulatory, nervous, or digestive system. The injured portions are affected by cancer and secrete hormones that cause changes in body structure [2]. In general, tumors are of two types: malignant or benign. Malignant tumors spread quickly into surrounding tissues and destroy them. It causes the cell to grow into another cell, which leads to the emergence of other tumors. Therefore, malignant tumors are inherently dangerous and pose a threat to human life. While benign tumors do not cause harm over time if they grow a lot, they pose a risk or may become malignant tumors [3]. The cancer has several symptoms, as extraordinary bleeding, swelling, weight lose extremely, and others, so the symptoms must be studied and known properly to provide appropriate treatment for patients, this necessitates a high-quality classification and prognostic system for tumor classification to malignant and benign.

There are more than 100 types of cancer that pose a threat to human life. Breast cancer is more common and more effective than other types of cancer. The danger elements of breast cancer are lack of physical practice, sex, obesity, liquor consumption, ionizing radiation, hormonal imbalance, premenstrual period, and advanced age. These factors above are not the only common factors, there may be other causes that lead to breast cancer, its aggressiveness, spread, or genetic makeup [4]. Research presented in the medical field related to cancer is individual of the main appealing topics. To accurately diagnose the disease and provide treatment for those suffering from cancer, accurate prediction systems must be available to detect tumors and cancers. Previous procedures were manual and clinical and not without errors.

Having systems that detect cancerous diseases at a precocious phase is significant in receiving timely treatment and prevention. Therefore, this will raise the stay-alive rates of persons with cancer [4]. Due to the development of the

*Corresponding author: oqbah_salim@tu.edu.iq

disease, it has become difficult for a doctor to diagnose cancer directly, or the disease may be diagnosed incorrectly. Therefore, there is a need for more research to discover and diagnose cancer in the human body. Today, with the development that has occurred, utilizing the different models of machine learning (ML) become feasible. These technologies can classify cancer images or identify patient data and symptoms to predict the disease. There are many ML models utilized in research of cancer, that have helped in diagnosing and detecting cancer. Thus, machine learning techniques can perform many intelligent calculations to disclose and forecast cancer at a precocious phase if patient data is collected. To provide the necessary treatment. The survey conducted by PubMed showed that statistics related to research work on cancer detection showed that many research studies have been published on cancer detection using machine learning to identify the tumor or cancer [4]. Machine learning was only useful for identifying or detecting cancer, but newer techniques in machine learning have mainly focused on diagnosing and predicting disease. Data mining techniques in machine learning are majority widely utilized in gene data expression and cancer classification. ML models are applied in the detection of types of cancer, as Fig. 1.

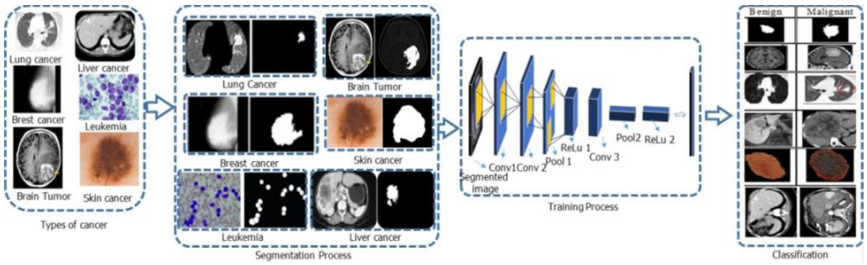

FIGURE 1. - Display ML in cancer detection [5]

2. MACHINE LEARNING TECHNIQUES

The machine learning phrase was first formulated in 1959 from Arthur Samuel and was utilized in programming electronic games [6]. Machine science is based on programming computer programs to perform various tasks and implement the instructions and commands assigned to them. The tasks rely on the data available to it and analyze it completely without human intervention [7]. Machine learning learns and trains statistical strategies and models to perform many functions without explicit instructions. Machine learning works on the principle that machines learn automatically without being explicitly programmed. ML is interested in the evolution of computer techniques to data arrival through training and learning [8]. Machine learning has been invested in many fields such as healthcare, commerce, social data, etc. [9]. ML is a type of artificial intelligence and is important for discovering learning outcomes based on a data-sampling approach. There are two basic steps to the training and learning process [10]. The unknown dependencies of the model are identified through the provided data set. The new results of the model are the outputs of predicting whether the identified dependencies are known or not. There are many machine learning models in healthcare research including biomedicine. Machine learning uses different models and techniques to generalize biological data with broad dimensions of the dataset. Machine learning learns in many ways, as Fig. 2.

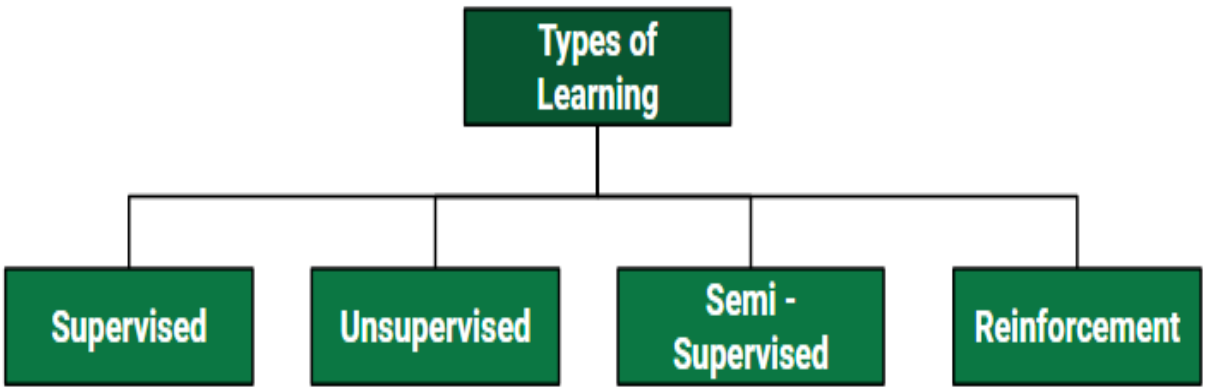

FIGURE 2. - Shows types learn in machine learning [11]

2.1 SUPERVISED LEARNING

This type of learning relies on categorical data. The techniques work on the principle of training. It requires learning the algorithm to summarize the nature of the hidden training information acceptably [12]. Thus, using this trained model to predict of data chosen. Data should be segmented into two portions: training and testing sets. The classifier is trained on the training group first, and the model's examination on the test group. Model power can be measured through performance gauges [13]. It can solve problems in classification and regression. The categorical values are separate in the classification process. While in the regression process, the values are continuous. Models in the classification are used to predict which sample belongs to which class. In regression, models are utilized to forecast the dependence of continuous data [14].

2.2 UNSUPERVISED LEARNING

Unsupervised learning uses datasets that include unlabeled inputs. The most common model in this learning is cluster analysis, which is used to combine information and explore hidden patterns or information in groups. Unsupervised learning requires training data except that the target value must be unknown. The model tries to group similar types of data through the discovery of the invisible types. The goal of learning is to discover types. The power of the classifier in this learning kind cannot be assessed because the value of the label is unidentified [14].

2.3 SEMI-SUPERVISED LEARNING

It operates on categorical value and non-categorical. It is a mixture of a supervised learning system and an unsupervised system that relies on classified and unclassified data. So the data must be unclassified longer length compared to named data. In this learning, a big change in performance power is observed when each of the labeled and unlabeled values is present. That is utilized to identify outliers. To classify data with high accuracy, categorical and non-categorical data sets must be used together [15].

2.4 REINFORCEMENT LEARNING

In this kind of learning the new model learns the pattern of acceptable behavior based on environmental awareness. Every action affects the environment to generate inputs that help learning models. Where it is a model that works to develop its performance based on the environment through interaction without human intervention [16].

2.4.1 K-NEAREST NEIGHBOR(KNN)

It is a model utilized to category data where a new state is created that is added to the current state within a specific region and the nearest neighbor with the same features similar to the state is calculated. k is a user-fixed value, where each particular case finds similar features present in the new case, and this leads to the new case being assigned to a similar category. [17- 20].

2.4.2 SUPPORT VECTOR MACHINE (SVM)

It is utilized to apply regression analysis and categorical data classification to a data set. It works on large groups of entries data. This algorithm partitions the dataset into a couple of categories, where classification is done depending on categorical data. It works by measuring the farthest margin among the super level and the data point closest to any categorical set of datasets. [4] [17-19] [21-28].

2.4.3 DECISION TREE

It is utilized in both regression and classification, because of its ability to work in both continuous and categorical variables. The decision tree looks such as a tree, at the top begins with a taproot node, and then sections into several branches to form a number of solutions. The model provides a solution and a documented process for each step to reach the goal is selected. The possible outcome of the given features and more specific solutions can be configured. The model is done to predict the indicated solutions to test the results in the prediction phase. [19] [22] [24] [29,30].

2.4.4 NAÏVE BAYES CLASSIFIER

It is a statistical model which depends on probability. These classifiers are used for large data sets and can make filtering decisions based on data points from various attributes and categories. The goal of this algorithm is to forecast the category label for a given set. This theory describes Bayesian classifiers assuming independence between predictors. [19][26].

2.4.5 RANDOM FOREST

It algorithm operates in the same approach as a decision tree. It produces several decision trees to form a forest. It reconstructed the trees by randomly selecting variables and data. To enable this algorithm, an ensemble classifier is used, which selects a group of features to develop every decision tree from the different decision trees. The final decision is made by the maximum votes generated by the different trees. [26] [31].

2.4.6 K-MEAN ALGORITHM

Clustering is a trending unsupervised ML model [32], which is utilized to divide a data set into a pre-determined fixed number k, which indicates several clusters for the recurrence approach. The model K-Mean clustering is clustering web searches through identifying similarities and checking the linkage rate of seeking outputs on any seek site such as yahoo, google, etc.

3. LITERATURE SURVEY

In the paper [4] the authors applied a number of SVM ensemble and SVM classifiers to breast cancer datasets. In addition to using RBF kernel and boosting method based on SVM to obtain high accuracy prediction as well as calculate evaluation parameters such as F-measure and ROC curve to build a robust model for training data. They found that the RBF kernel and boosting method based on SVM gives higher accuracy than other models.

In this paper [17] the authors found there are several classifiers and features to identify genes extracted from fine-grained corpora that include a lot of noise. Three data sets were selected: colon cancer, leukemia, and lymphoma as well, which includes approximately 72, 62, and 47 samples. They used Euclidean distance, Spearman, and Pearson correlation coefficients, information exchange, percentage of the signal to the noise, and information gain for the characteristic pick. They utilized MLP, K-NN, SVM, and SOM for classification, the empirical outcomes achieved on every dataset used with each classifier described above showed that the best result was in the leukemia dataset with an accuracy of 97.1%.

In the paper [18], the researcher used a new way depending on the feature choice approach of classification highest dimensions in cancer microarray value. In this method, filtering techniques were used to improve PSO and the percentage of the signal to the noise. They found that PSO offers the best results if used with k-NN, SVM, and PNN. The leukemia dataset used includes 72 samples, which include (7129) genes. The colon cancer includes (62) cases, which contain (2000) genes. Breast cancer included 97 cases containing 24,481 genes. They found that PSO with other algorithms gave results with 100% accuracy on a breast cancer dataset.

In the study [19] compares various models in ML: and SVM, kNN, NB, and C4.5 at this same dataset found on the WBCD website that has 699 samples and 11 attributes with integer values. The SVM algorithm provides a greater accuracy of 97.13% among all algorithms, with less error rate than was applied in the data mining models on the WEKA website.

In this paper [23] a data mining technique was applied to a wide range of data to find important knowledge. To reduce the set of existing attributes and to find the data dependence in the dataset, approximate set theory was used. To improve the feature selection for the ovarian cancer dataset at different stages the hybrid particle-optimized genetic swarm was used. A multilayer SVM model was chosen to classify different and normal stages at ovarian cancer utilizing enhanced features. The datasets were obtained of the ovarian cancer website, which includes more than 493 samples with 12,042 genes. The classifier ANN, and Naïve Bayes Multiclass SVM were with experimental results of accuracy 93%, 90%, and 96%, respectively.

The paper [26] used three machine learning algorithms: RF, SVM, and Naïve Bayes, the result at RF was the greater suitable, it produced an accuracy of about 99.42%, and the results of NB and SVM were 98.24% and 98.8%, respectively.

In the paper [33] the authors used the PSO model to classify and predict the survival of patients by using data on gene expression. The PSO model reduces dimensions by applying the probability NN. Experimental results achieved by applying the model to a B-cell lymphoma dataset of more than 240 samples were effectively up to 80% accurate in predicting patients' survival.

In the paper [34] the authors tried to solve the impression of breast cancer syndrome datasets that have improved and evolved through social media and seeking studies utilizing the K-medoid combinations. optimized K-medoid combinations have also been developed that help improve assembly performance and efficiency by remapping breast cancer syndrome images and displaying the negative medium silhouette to another group after the primary K-medoid grouping process.

In the paper [35] the researchers worked to find variously expressed genes among the advanced and early status of various cancers by using data from RNA series. The significance of these samples is tested extensively by improving prediction algorithms using linear discriminant analysis and nearest neighbor algorithms. The results in this paper show that analysis of cancer, in general, may be similar to criterion analyses of single cancers to characterize the biologically associated DE genes. It would help improve predictive algorithms' strength in evolving cancer detection methods. Genetic microarray data usually contain a large multiple of genes benchmarked to the smaller number of

existing tests. This task can thus be motivated to detect a minor sub-group of genes related to data of microarray gene, the selected chromosome can be utilized exclusively to sequence the cancer subdomain with high accuracy.

In the paper [36] the researcher presented several classification and prediction ways and attribute selection ways of genes annotated in microarray data. Researchers were able to explore the competence of different classification models such as beam basis function, SVM, DT, multi-class perceptron, and RF. Validation of 10 fold was implemented to accuracy measure of the algorithm including K-means. In addition, the attribute selection ways were gauged by correlation attribute selection, Chi-Square, and SVM-RFE, in the end, the researchers obtained the most effective result of feature selection by using SVM-RFE in identifying the genes of interest with 100% accuracy.

In this study [37], a new hybrid system that depends on neural networks and association base mining is implemented. Which adopts evolutionary techniques where the amplitude problem was used to find breast cancer. The arm improvement relies on the Grammatical evolution to identify the most important features and reduce the measurement capacity to determine the correlation between SNPs, after which the NN model is used for effective distribution.

The paper [38] uses a new hybrid model of the gray wolf optimizer that is combined with a decision tree. It is designed as a classifier to select the minimum important genes from among many genes to identify cancer. The model and other well-known algorithms such as SOM, BPNN, PSOC4.S, C4.S, and SVM applied the test samples to 10 genes. The results on two datasets show that the applied model is better than other models.

This paper [39] designed a framework to classify cancer sites by studying bodily mutations utilizing ML techniques. Researchers extracted information regarding patients, genes linked to the mutation, and the linked mutation chromosomes from the somatic mutations catalog cancer database, also combined gene function and linked mutation genes utilizing gene tracks in the database of the genomes and genes.

This study [40] designed a high-level distributed method for monitoring typical and abnormal MRI brain images. Analysis of brain tumors depends on the radiological existence and syndrome. Scanning with magnetic resonance is more importantly means for detecting brain tumors. Currently, there are various methods. It has been used to image brain cancer. But for the most effective examination, the image is pre-processed and the image features are extracted. The advanced classification is implemented for brain cancer. When distinguished ML models: KNN, SVM, and combined classifier as SVM-KNN are utilized to request 50 pictures. It is clear from the outcomes that the SVM-KNN combined classifier showed the best ranking with a 98% accuracy rate among others.

The paper [41] proposed the use of ML models to classify liver cancer utilizing a combined dataset in magnetic resonance imaging and two-dimensional computed tomography. For classification, 10 improved features and hybrids were identified by employing feature selection methods. ML techniques were used: SVM, random forest, J48, and Multilayer Perceptron (MLP) with 10 points of double cross-validation approaches. MLP appeared an overall accuracy of 97.44% in CT and 95.78% in MRI. An enhanced hybrid features dataset was created. The accuracy of MLP was 99% among all other techniques.

In this study [42], it was proposed to do statistical hypothesis checks like t-test and mann wilcoxon whitney test utilizing ML models like neural networks, decision trees, KNN, etc. To identify the more important genes in detecting colon cancer patients. The dataset is normalized in two stages so that this is more efficient and effective than other methods. The dataset dimensions were decreased utilizing principal components analysis (PCA). It extracted 20 efficient genes in each group that could be useful for precocious prediction of colon cancer.

In the paper [43]. An efficient and accurate computational method for selecting gene identity has been established. Initially, the t-test approach is A traditional technique to reduce the scaling process of a data set. and then; This method has been used to detect the gene code of interest through swarm optimization of selected particles, the specific strategy that contains (SRBCT) data to initialize subsections namely non-Hodgkin's lymphoma, rhabdomyosarcoma, (ASW) syndrome, Ewing sarcoma. to other clusters after the primitive K-medoid clusters.

4. COMPARISON OF MACHINE LEARNING

Through the reviews, we discovered that different types of cancer are diagnosed and predicted differently. Characteristics and behavior vary depending on the type of cancer. Ovarian cancer occurs only in females. ML strategies have been implemented to an ovarian cancer dataset with mixed results. Liver cancer occurs in the liver, and ML models have been used to detect the disease with high accuracy. Colon cancer begins in the rectum or colon. ML models are implemented to analyze and detect colon cancer. Datasets were collected from many sources and ML strategies were executed to predict and diagnose the illness, and the results were highly accurate.

We have classified cancer types as shown in the table below, table 1 presents the classification models implemented in several research studies to diagnose and predict cancer. This table also displays the samples contained in the data set and approved for acquisition on accuracy. The observation was made on the basis of a review of papers using machine learning models for early cancer diagnosis and prediction. Before applying ML models, the dataset should be gathered and standardized. After that, the unimportant data is processed for classification and diagnosis. Data dimensions' decrease because it is not possible to use the entire data in the dataset.

Table 1.- Show accuracy measured of ML techniques used for cancer

Ref	ML Technique	Sample	Accuracy	Time learning	Type of cancer
[15]	SVM-RFE	39	100		Ovarian Cancer
[23]	SVM ANN Naïve Bayes	498	96% 93% 90%		Ovarian Cancer
[32]	K-Means	155	97%		Ovarian Cancer
[33]	SVM	156	86.67%		Liver Cancer
[41]	SVM random forest J48 Multilayer Perceptron (MLP)	254	95.78% 94.44% 94.44% 95.88%	0.31 0.11 0.16 0.42	Liver Cancer
[44]	Particle Swarm Optimization	-	93.3%	42.429959	Liver Cancer
[17]	SASOM		93.6%		Colon Cancer
[20]	KNN using Biogeography-Based Optimization		80%		Colon Cancer
[42]	Quadratic Discriminant Analysis Naive Bayes SVM Linear Kernel Logistic Regression Linear Discriminant Analysis Ada-Boost KNN Neural Network Decision Tree	461	90.99 % 83.26 % 94.77 % 96.03 % 94.97 % 96.66 % 97.49 % 97.08% 100%		Colon Cancer
[45]	recursive feature elimination and SVM with fuzzy granular (FGSVM-RFE)	62	100%		Colon Cancer
[18]	PSO-KNN PSO-SVM	97	100%		Breast Cancer
[19]	SVM C4.5 NB k-NN	699	97.13% 95.13 % 95.99% 95.27%		Breast Cancer

[24]	Decision Tree and Support Vector Machines	699	91%		Breast Cancer
[26]	Random Forest SVM Naïve Bayes	699	99.24% 98.8% 98.24%		Breast Cancer
[31]	Random Forest (RF)	699	99.8%	13.05	Breast Cancer
[30]	Decision Tree (C4.5) ANN SVM	1189	93.6% 94.7% 95.7%		Breast Cancer
[46]	Naïve Bayes (NGB) K-nearest neighbor (kNN) Support Vector Machines (SVM) Decision trees (DT)	1102	87% 88% 90% 91%		Breast Cancer
[47]	Decision trees (DT) SVM NB		100% 98.42% 69.60%		Breast Cancer

5. CONCLUSION

The review presents effective techniques for cancer classification and prediction through research in the field of bioinformatics. We infer from this review that several automatic cancer prediction strategies depend on machine learning models containing clustering and classification techniques. This study provided a wide review of the variety of ML concepts and classification ways used to predict and detect cancer in different types of cancer datasets, like ovarian cancer, liver, colon, and breast cancer.

The results achieved by several authors to overcome challenges using several machine learning techniques and applied to datasets with different number samples are presented in detail and tabulated. The most successful and superior models are SVM, random forests, and decision trees, which give the highest accuracy. Their superiority was noted when the number of samples was large or small, meaning they worked efficiently if the data sets were large or small. But, in the case of large data sets, the accuracy decreases slightly compared to small data sets. It is necessary to use dimensionality reduction models on large data sets to identify the most important features. There are still possibilities for improving and developing early-stage cancer prediction. There are many data sets available to reveal more knowledge about cancer.

FUNDING

None

ACKNOWLEDGEMENT

The authors express their extreme thanks and gratitude to the researchers for the resources they provided and appreciate the suggestions from the reviewers.

CONFLICTS OF INTEREST

The authors declare no conflict of interest

REFERENCES

[1] World Health Organization (WHO), "Cancer Fact Sheet," available: http://www.who.int/mediacentre/factsheets/fs297/en/, Oct. 2017.

[2] "Cancer and oncology," Medical News Today, updated Jan. 24, 2024. [Online]. Available: http://www.medicalnewstoday.com/info/cancer-oncology

[3] G. Cooper, The Cell: A Molecular Approach, 2nd ed. Sunderland, MA: Sinauer Associates, 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9963/
[4] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "SVM and SVM ensemble in breast cancer prediction," PLOS One, vol. 12, no. 1, pp. 1-16, Jan. 2017, doi: 10.1371/journal.pone.0161501.
[5] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons, and challenges," J. Infect. Public Health, vol. 13, no. 9, pp. 1274-1289, 2020.
[6] A. Boutorh and A. Guessoum, "Classification of SNPs for breast cancer diagnosis using neural-network-based association rules," in 12th Int. Symp. Programming and Systems (ISPS), IEEE, 2015.
[7] O. S. Atiyah et al., "A comparison of COVID-19 cases classification based on machine learning approaches," Iraqi J. Electr. Electron. Eng., vol. 18, no. 1, pp. 139-143, 2022.
[8] Y.-X. Wang and M. Hebert, "Learning to learn: Model regression networks for easy small learning," in European Conf. Computer Vision (ECCV 2016) Lecture Notes in Computer Science, vol. 9910, pp. 616-634, 2016, doi: 10.1007/978-3-319-46466-4_37.
[9] O. S. Atiyah et al., "Evaluation of COVID-19 cases based on classification algorithms in machine learning," Webology, vol. 19, no. 1, pp. 4878-4887, 2022.
[10] E. Alpaydin, Introduction to Machine Learning. Cambridge, MA: MIT Press, 2009.
[11] "Supervised machine learning," GeeksforGeeks, available: https://www.geeksforgeeks.org/supervised-machine-learning/, Feb. 27, 2024.
[12] N. S. Abd, O. S. Atiyah, M. T. Ahmed, and A. Bakhit, "Digital marketing data classification by using machine learning algorithms," Iraqi J. Electr. Electron. Eng., vol. 20, no. 1, 2024.
[13] D. Jayashree, S. K. Dash, S. Dash, and M. Swain, "A classification technique for microarray gene expression data using PSO-FLANN," Int. J. Comput. Sci. Eng., vol. 4, no. 9, pp. 1534-1539, Sep. 2012.
[14] S. Akanksha and P. Kaur, "Optimized liver tumor detection and segmentation using neural network," Int. J. Recent Technol. Eng. (IJRTE), vol. 2, no. 5, pp. 7-10, Nov. 2013.
[15] Y. Hu et al., "A comparison of neural network and fuzzy c-means methods in bladder cancer cell classification," in Proc. 1994 IEEE Int. Conf. Neural Networks (ICNN'94), 1994, pp. 3461-3466, doi: 10.1109/ICNN.1994.374891.
[16] T. O. Ayodele, "Types of machine learning algorithms," New Adv. Mach. Learn., vol. 3, pp. 19-48, 2010.
[17] S. B. Cho and H. H. Won, "Machine learning in DNA microarray analysis for cancer classification," in Proc. 1st Asia-Pacific Bioinformatics Conf. Bioinformatics, vol. 19, pp. 189-198, Jan. 2003.
[18] S. Barnali and D. Mishra, "A novel feature selection algorithm using particle swarm optimization for cancer microarray data," in Int. Conf. Modeling Optim. Comput. (ICMOC-2012), ELSEVIER Procedia Eng., vol. 38, pp. 27–31, 2012.
[19] A. Hiba, H. Mousannif, H. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," in 6th Int. Symp. Front. Ambient Mobile Syst. (FAMS 2016), ELSEVIER, vol. 83, pp. 1064–1066, 2016.
[20] P. Ammu, K. C. Siva Kumar, and M. Sathish, "A BBO-based feature selection method for DNA microarray," Int. J. Res. Stud. Biosci. (IJRSB), vol. 3, no. 1, pp. 201-204, Jan. 2015.
[21] P. Ammu and P. V. Preeja, "Review on feature selection techniques of DNA microarray data," Int. J. Comput. Appl., vol. 61, no. 12, pp. 39-44, Jan. 2013.
[22] P. Ramachandran, N. Girija, and T. Bhuvaneswari, "Early detection and prevention of cancer using data mining techniques," Int. J. Comput. Appl., vol. 97, no. 13, pp. 48-53, Jul. 2014.
[23] P. Yasodha and N. R. Anathanarayanan, "Analyzing big data to build knowledge-based system for early detection of ovarian cancer," Indian J. Sci. Technol., vol. 8, no. 14, Jul. 2015, doi: 10.17485/ijst/2015/v8i14/65745.
[24] K. Sivakami, "Mining big data: Breast cancer prediction using DT-SVM hybrid model," Int. J. Sci. Eng. Appl. Sci. (IJSEAS), vol. 1, no. 5, pp. 418-429, Aug. 2015.
[25] P. Rajeswari and G. S. Reena, "Human liver cancer classification using microarray gene expression data," Int. J. Comput. Appl., vol. 34, no. 6, pp. 25-37, Nov. 2011.
[26] B. Dai, R. C. Chen, S. Z. Zhu, and W. W. Zhang, "Using random forest algorithm for breast cancer diagnosis," in 2018 Int. Symp. Comput., Consum., Control (IS3C), IEEE, pp. 449-452, Dec. 2018.
[27] D. Urun, T. Glasmachers, and C. Igel, "A unified view on multi-class support vector classification," J. Mach. Learn. Res., vol. 17, pp. 1-32, 2016.
[28] A. Mennatallah, M. Goldstein, and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection," in Proc. ACM SIGKDD 2013 Workshop on Outlier Detection and Description, pp. 8-15, 2013.
[29] C. Vikas and S. Pal, "Data mining techniques: To predict and resolve breast cancer survivability," Int. J. Comput. Sci. Mobile Comput., vol. 3, no. 1, pp. 10-22, Jan. 2014.

[30] L. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. R. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence," J. Health Med. Inform., vol. 4, no. 2, 2013, doi: 10.4172/2157-7420.1000124.

[31] N. Cuong, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," J. Biomed. Sci. Eng., vol. 6, pp. 551-560, 2013, doi: 10.4236/jbise.2013.65070.

[32] T. Arunanand, A. Nazeer, M. J. P., and M. Pradhant, "A nature-inspired hybrid fuzzy c-means algorithm for better clustering of biological data sets," in IEEE Int. Conf. Data Sci. Eng. (ICDSE'14), pp. 76-82, 2014.

[33] X. Rui, X. Cai, and D. C. Wunsch II, "Gene expression data for DLBCL cancer survival prediction with a combination of machine learning technologies," in Proc. 2005 IEEE Eng. Med. Biol. 27th Annu. Conf., pp. 894-897, 2005.

[34] Q. Ping, C. C. Yang, S. A. Marshall, N. E. Avis, and E. H. Ip, "Breast cancer symptom clusters derived from social media and research study data using improved K-Medoid clustering," IEEE Trans. Comput. Soc. Syst., vol. 3, no. 2, pp. 63-74, Jun. 2016.

[35] S. Mishra, C. D. Kaddi, and M. D. Wang, "Pan-cancer analysis for studying cancer stage using protein and gene expression data," in 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), pp. 2440-2443, 2016.

[36] P. Mehdi, J. Y. Yang, M. Q. Yang, and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data," BMC Genom., Int. Conf. Bioinformatics Comput. Biol. (BIOCOMP'07), pp. 25-28, Jun. 2007, doi: 10.1186/1471-2164-9-S1-S13.

[37] B. Aicha and A. G., "Classification of SNPs for breast cancer diagnosis using neural-network-based association rules," in 12th Int. Symp. Program. Syst. (ISPS), IEEE, 2015.

[38] M. Vosooghifard and H. Ebrahimpour, "Applying grey wolf optimizer-based decision tree classifier for cancer classification on gene expression data," in 5th Int. Conf. Comput. Knowl. Eng. (ICCKE), Mashhad, pp. 147-151, 2015.

[39] Y. Chen et al., "Classification of cancer primary sites using machine learning and somatic mutations," BioMed Res. Int., 2015.

[40] K. Machhale, H. B. Nandpuru, V. Kapur, and L. Kosta, "MRI brain cancer classification using hybrid classifier (SVM-KNN)," in Int. Conf. Ind. Instrum. Control, Pune, pp. 60-65, 2015.

[41] S. Naeem et al., "Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images," Appl. Sci., vol. 10, no. 9, p. 3134, 2020, doi: 10.3390/app10093134.

[42] M. Fahami et al., "Detection of effective genes in colon cancer: A machine learning approach," Informatics Med. Unlocked, vol. 24, p. 100605, 2021, doi: 10.1016/j.imu.2021.100605.

[43] K. Subhajit et al., "A particle swarm optimization based gene identification technique for classification of cancer subgroups," in 2nd IEEE Int. Conf. Control, Instrum., Energy Commun. (CIEC), 2016.

[44] S. Akanksha et al., "Optimized liver tumor detection and segmentation using neural network," Int. J. Recent Technol. Eng. (IJRTE), vol. 2, no. 5, pp. 7-10, Nov. 2013.

[45] T. Yuchun et al., "Recursive fuzzy granulation for gene subsets extraction and cancer classification," IEEE Trans. Inf. Technol. Biomed., vol. 12, no. 6, pp. 723-730, Nov. 2008.

[46] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," J. Pers. Med., vol. 11, no. 2, p. 61, 2021.

[47] A. Alhasani et al., "A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining," Methods, vol. 7, no. 9, 2023.