Review of Machine Learning Algorithm on Cancer Classification for Cancer Prediction and Detection

Arvind Jaiswal
Research Scholar(CSE),
Bhagwant University, Ajmer(Rajasthan), India
arvindjsir@gmail.com

Dr. Rajeev Kumar
Associate Professor, (CS&IT),
Teerthanker Mahaveer University, Moradabad, (U.P.), India
drkiit1985@gmail.com

***Abstract*— Cancer is a critical disease from many years. This leads to death if it is not diagnosed at early stage. It is a topic of concern because actual treatment of this disease is not found till date. Patients having this disease can only be saved if and only if it is found in early stage (I and II). If it is detected in latter stage (III and IV) then chance of survival is very less. Machine learning and data mining technique is very helpful technique to handle this problem. Machine learning is demonstrating the promise of producing consistently accurate estimates. Machine learning system effectively "learns" how to estimate from training dataset of completed operations. There are various techniques available in Machine Learning to predict the cancer on the basis of collected standard datasets. The datasets may have been recorded by social media, healthcare websites and some other repositories. We need to apply some classifiers of Machine Learning Techniques on these dataset to detect the cancer in a human. The main aim of the review is to help the research on accurate estimation, i.e. to ease other researchers for relevant correct estimation studies using machine-learning techniques. Our review suggests that these techniques are competitive with traditional estimators on datasets and also demonstrate that these methods are sensitive to the data on which they are trained.**

***Keywords*— Machine Learning, Cancer, datasets, Machine Learning Techniques, data mining technique.**

I. INTRODUCTION

World Health Organization (WHO) has reported that cancer is the world's second biggest killer after ischaemic heart disease and stroke [1]. Cancer is a group of diseases. It is a dangerous disease that characterized by the nature of the cell inside the body which has no control. It involves abnormal growth of cells. It spreads and affects very fast to other parts of body. Cancer damages the human body gradually when cells starts growing uncontrollably to form many lumps of tissue inside the human body called tumours. It is not necessary that all kind of tumours are cancerous. Some kind of tumours is not spread in the body. tumours may grow and interact with the other parts of the body. That part may be nervous system, digestive system or circulatory system. The effect of infected parts of the body releases the hormones that cause change in the body [27]. Tumours are basically of two types i.e. benign or malignant. Malignant is spread into the surrounding tissue. Cell can grow to other cell and destroy the surrounding tissue that causes other tumour to develop. So malignant tumour can be a life-threatening and more dangerous in nature. Benign tumour usually do not cause much damage but can become more dangerous if they grow a lot or they might become malignant after certain amount of time[28].

Cancer has various symptoms such as tumour, abnormal bleeding, more weight loss etc. To provide appropriate treatment to the patients, symptoms must be studied properly and an automatic prediction system is required which will classify the tumour into benign or malignant. There are near about 100 types of cancers affecting human body.

Among all the types of cancer, the very popular and influenced is Breast cancer. The main risk factors of breast cancer include sex, obesity, less physical exercise, intakes of alcohol, hormonal misbalance during menopause, ionizing radiation, pre-menstruation, children at later age or not at all, and older age. The above factors are not the common factors. There may be some another reason to cause breast cancer with different stages or spread, aggressiveness and genetic makeup [30].

In medical field, research on cancer is one of the challenging, attractive and main points of focused area. To provide the appropriate treatment to the patients, there is a need for accurate automatic prediction systems for tumour and cancers. Previous treatments are manual and clinical based. Such conventional classification methods have some limitations such as:

- Slow diagnostic process.
- Tumor/cancer prediction based on pathological reports, which may increase the patient's efficacy.
- For cancer classification and prediction, requires various types of clinical courses.

As per the above discussion, it would be nice to have such a system that would allow detecting and preventing the cancer at an early stage. This can increase the survival rates for those who are going to effect of cancer [30]. To prognosis and diagnoses cancer by a physician may become difficult by seeing the human bodies until their cells are treated. Further a research is required to diagnose and prognosis the cancer in a human body.

In today's world various machine learning approaches are utilized. Such systems can detect and classify cancers images or patients data like age and symptoms. Machine learning is not new in the field of research on cancer. Artificial neural networks (ANNs) and decision trees (DTs) are already preowned in these sectors since last 20 years for detection and diagnosis of cancer. Hence Machine Learning can implement many computational intelligent techniques for the prediction of cancer at early stage if data of patient are collected. In the view of providing better treatment to the patient, it is important to precisely predict different type of tumours.

The survey made by latest PubMed provides statistics about research work on cancer detection methods. It shows that lots of papers are published, which are based on the relation between cancer and machine learning. Maximum papers point of interest is the use of machine learning methods for identification, classification, detection of cancer or tumour. In past machine learning is only useful for cancer detection and diagnosis, but recent systems mainly focused on prediction and prognosis of cancer. Among machine learning, data mining algorithms are the most commonly used, for classification of gene expression data and cancer.

II. MACHINE LEARNING TECHNIQUES

Machine Learning was originated by Samuel in 1950 to play strategic games like chess. It is the mechanism of making machines to learn automatically without being explicitly programmed. The main focus of Machine Learning is to develop a computer program which can access the data and use this data for learning purpose. It is the ability of machine to make use of statistical techniques and advanced algorithms to make more powerful prediction and making the data driven system more powerful by replacing the rule-based system. Machine Learning can be used in many fields such as finance, retail, health care and social data [3].

Machine Learning (ML) is a subdivision of AI. ML is useful in order to infer the learning outcome on the basis of behavior of data samples. There are mainly two phases of learning process [31]: (i) On the basis of dataset provided, the unknown dependencies are to be estimated for the system and (ii) New output of the system is to predict if estimated dependencies are known.

There are many applications in the area of biomedical research where ML fits suitably. ML uses different techniques and algorithm to generalize the biological sample of n-dimensional spaces for a given set of datasets.
There are many types of Machine Learning Techniques.

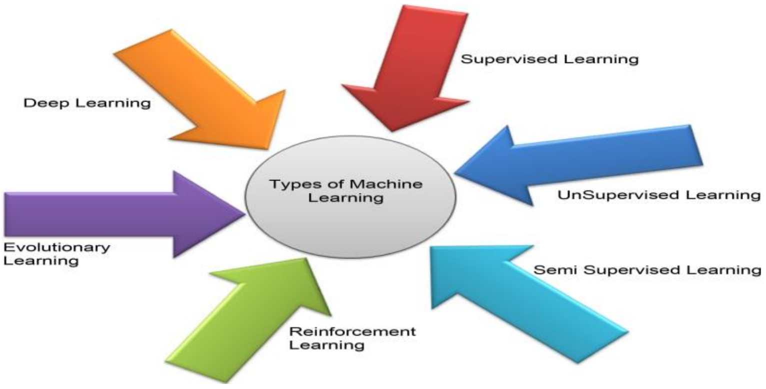

Fig. 1 Types of machine learning techniques

Supervised Learning (Classification Approach)

Supervised learning involves training the model on the labeled data and uses this trained model to make predictions on the new data. It involves splitting of data into two sets including training set and testing set. First the model is trained on training set and afterwards the performance is tested on the testing set. The performance of the model can be evaluated using performance metrics [4]. Supervised learning can be classification problem or regression problem. In supervised classification, the labeled value is a discrete value. The algorithms in this are used to classify to which class or category the problem belongs. On the other side, the models are used to predict the outcome based on continuous (numeric) data is supervised regression learning [4]. For the classification of raw data, first the data is selected and then preprocessing is performed in which all NA values are removed. Then the data is normalized using z-score or min max normalization. Once the normalization is performed feature selection procedure is applied to select the best features. After the features are selected, some supervised ML algorithms includes K Nearest neighbor, Decision trees, Support Vector Machines, Naïve Bays Classifier, Neural Network and Ensemble methods [3] are used for classification of raw data. The labeled training data is used to estimate to the desired output. e.g. ANN, Decision Tree, Random Forest, SVM, kNN, Gaussian Process regression, Naïve Bayes Classifier, Max Entropy classifier.

Unsupervised Learning (Clustering Approach)

Unsupervised Learning also involves training of the data except for the fact that the labeled value or target value is not known. In this, machine try to cluster the similar type of the data by finding the hidden pattern. Rather than making prediction, the main aim of unsupervised learning is to discover the patterns. The performance of the model in unsupervised learning cannot be evaluated as the label value is absent or unknown. The algorithms involved in unsupervised learning are K-mean clustering, Association Rule Mining, Topic Modelling and Dimensionality Reduction Techniques [3]. The notion of the output during the learning process is not known. e.g. K-means. PCA, Latent variable model, Hebbian learning.

Semi-Supervised Learning

As supervised learning works on labeled data and unsupervised learning on unlabeled data, then a lot of information is lost from labeled data which can be obtained from unlabeled data. So, in this case semi-supervised learning comes to mind. It is a mixture of supervised and unsupervised learning in which it takes both the unlabeled and labeled data. Labeled data should be of shorter length as compared to unlabeled data. The idea behind semi-supervised learning is that there is a considerable change in performance when both labeled and unlabeled data is used in conjunction. The training set used is of shorter length. It is normally used to detect outliers.

The datasets are, labeled and un-labeled, used to classify the data in better way. e.g. Self-training [34], Generative Models [34], S3VM, Graph-based method, Multi-view Learning, Mixture model

Reinforcement learning

Reinforcement Learning works by developing a system which improves its performance by taking feedback from the environment and taking possible steps to improve them. It is an act of learning from environment by interacting with it without any help from humans. It is an iterative process.

Evolutionary Learning

This biological evolution learning can be considered as a learning process: biological organisms are adapted to make progress in their survival rates and chance of having off springs. By using the idea of fitness, to check how accurate the solution is, we can use this model in a computer [5].It is an inductive process of self-learning based on previous experience. e.g. Partitioning transduction, Agglomerative transduction, Manifold transduction

Deep learning

This branch of machine learning is based on set of algorithms. In data, these learning algorithms model high-level abstraction. It uses deep graph with various processing layer, made up of many linear and nonlinear transformation.

Based on training input data and output data, tries to predict the new output when a new input is induced. e.g. Model Regression Network [33].

From the above definitive terms, we are able to differentiate a mix of machine learning algorithms for different purposes. The algorithms are described in the following section in brief to know more.

A. Artificial Neural Network (ANN)

A mathematical model connected with neurons in the brain is used to process the information in biological nervous system for computation. Neural Networks work in a system which are adaptive and allows to change the structure of neurons in learning phase too. Relationship between neurons can be easily framed. They are able to ascertain the pattern and develop the respective clusters within the data moderately. The learning process in ANN facilitates to classify the data and categorize the pattern. The neurons can be arranged and connected to form layers within and between the layers is called network structure [6,7,8,9,10,11,14,15,16,20,21,36]. The categorization of ANN is done as below:

- Single layer FF network
- Multilayer FF network
- Single node with its own feedback
- Multilayer recurrent network

B. Decision Tree

A Decision Tree [16, 17, 20, 23, 25] is to find the possible solutions of a given problem based on certain conditions that takes decision and the solutions are presented graphically for better understanding. The format of decision tree just like a tree starts with a root node at the top and then several branches are grown into a number of possible solutions. DT are useful to make decisions of the process. It gives a systematic solution and documentation process at each step. We can select a possible outcome from known alternatives and can identify more potential solutions.

C. Random Forest

A Random Forest Algorithm [13, 26] takes the decision tree concept a step further by producing a big number of decision trees to make a forest. These trees are reformed on the basis of selection of data and variables randomly. To enable this model, another classifier is used called ensemble classifier which starts by identifying a key set of features to grow each decision tree. It is a decision from multiple decision trees. The maximum votes come from many trees lead to the final decision.

D. Support Vector Machine

SVM [7,9,11,14-17,21,23,25,26,30,38,39] is a concept that is used to classify the labeled data and to apply regression analysis on the given dataset. There is a hyper plane to have sets of input data where SVM divides the dataset into two classes as the classification is done on the basis of labeled data in the best possible way. The distant margin is measured between the hyper plane and the nearest data point from either set of classified datasets

E. k-Nearest Neighbor (kNN)

kNN [7, 11, 22, 25] is a data classification algorithm that detects a new case to be with the existing case within a defined area by calculating nearest neighbor with similar features of the case. The value of k (where k is some user specified constant) would find all the similar existing features case with the new case and surround all the case so that it could be possible to identify the new case for the similar category.

F. Naïve Bayes Classifier

The Bayes' theorem describes the NB classifier [25, 26] with independence assumption between predictors. Bayesian classifiers are statistical classifiers based on probability. This is particularly used for large datasets where decision may be taken for filtration on the basis of data points of different class and attributes. The purpose of Bayesian theorem is to predict the class label for a given tuple.

G. K-means Algorithm

A very popular unsupervised ML technique is K-means [19] used for cluster analysis. Here k is a pre-defined constant that represents the number of clusters made for iterative method. The verities of K-Means clustering applications are to cluster the web pages by examining the similarities and identifying the relevance rate of the search results in any search engine like Google, yahoo etc.

III. LITERATURE SURVEY

In paper [2] author chased to resolve impression of syndrome clusters in breast cancer scraps evolved from both social media and research study data using improved K-medoid clustering. Also developed improved K-medoid clustering which helps to upgrade the clustering performance by reassigning some of the negative average silhouette width (ASW) syndrome to other clusters after initial K-medoid clustering.

In paper [7] author explored many features and classifiers to select extracted genes from microarray which have many noises. They have taken three datasets: Leukemia cancer, Colon cancer and Lymphoma cancer which has the sample 72, 62 and 47 respectively. They have used Pearson's and Spearman's correlation coefficients, Euclidean distance, information gain, mutual information and signal to noise ratio for feature selection. For classification, they used MLP, kNN, SVM and SOM. They performed experimental results with all the dataset given and shown the best result for accuracy is 97.1% on Leukemia dataset with all the classifier shown above.

In paper[8] author used PSO for the prediction of patient survival using gene expression data. PSO reduces the dimensionality by implementing Probabilistic NN. The experimental results of PSO/PNN on B-cell Lymphoma dataset of 240 sample was more effective up to 80% accuracy in survival prediction.

In paper [11] author proposed a novel approach based on feature selection method in order to classify high dimensional cancer microarray data. This approach uses one of the filtering techniques for optimization: signal-to- noise ratio (SNR) and PSO. They demonstrated that PSO gives better result when implementation is done along with SVM, k-NN and PNN. They have described the dataset of Leukemia having 72 instances with 7129 genes, Colon cancer having 62 instances with 2000 genes, DLBCL having 77 instances with 6817 genes and Breast cancer having 97 instances with 24481 genes. The accuracy they found PSO along with other classifiers gave 100% in Breast cancer case.

In paper [12] author seeked to extract differentially expressed (DE) genes between early and advanced cases of multiple cancer types through the use of RNA sequencing data. The importance of these genes is further examined by developing predictive models using K-nearest neighbor and linear discriminate analysis classifiers. The outcomes of the paper state that a pancancer analysis may be highly equivalent to standard analyses of individual cancers for describing biologically relevant DE genes and can assist in developing powerful predictive models for cancer progression. Microarray gene expression information normally consists of an enormous number of genes contrasted with less number of tests accessible. In this manner, it is the motivating assignment to recognize a little subgroup of pertinent genes from microarray gene expression information where the distinguished chromosome can exclusively be utilized for precisely arranging the cancer subspace. Consequently, In paper [24] a reckoning proficient but precise gene ID strategy has been nominated. At the commencement, the t-test technique is antiquated to diminish the measurement of the dataset and after that; the recommended particle swarm optimization based approach has been utilized to discover helpful genetic code. The adduced strategy has been connected on the small round blue cell tumor (SRBCT) information to arrange the four subdivisions particularly neuroblastoma, nonHodgkin lymphoma, rhabdomyosarcoma and Ewing sarcoma. (ASW) syndrome to other clusters after initial K-medoid clustering.

In paper [18] author studied many classification methods and feature selection methods for expressed genes in microarray data. They were able to find the efficiency of the various classification methods like: SVM, Radial Basic Function, Mult-Layer Perceptron, DT and RF. The 10-fold cross validation had been applied to calculate the accuracy of the classifier includes: K-means. Further the efficiency of the feature selection methods was measured by SVM-RFE, Chi-Squared and Correlation based feature selection (CFS). In the conclusion, the authors got the best efficient result by SVM-RFE feature selection methods with 100% accuracy to identify the significant genes.

In paper [21] applied data mining technique on large amount of data to discover the valuable knowledge. Rough set theory was utilized to find the data reliance and reduce the feature set contained in the data set. The Hybrid Particle Genetic Swarm Optimization is used to optimize the selected features of ovarian cancer at different stages. Multi class SVM is adopted to classify normal or different stage of ovarian cancer using optimized feature set. The datasets are taken from the TCGA portal (http:// tcga-data.Nci.nih.gov/) of ovarian cancer composed of 12042 genes with 493 instances. The classifier Multiclass SVM, ANN and Naïve Bayes analyzed their experimental result of accuracy 96%, 93% and 90% respectively. The other dataset was taken from NUH Singapore (http://www.nuh.com.sg/#) of blood test consist of 172 instances with 28 features. Their results were shown with three classifier SVM, ANN and Naïve Bayes as 98%, 95% and 93% accuracy respectively.

Paper [25] compared different ML algorithms: SVM, C4.5, NB and kNN for that dataset is available on WBCD which has 699 instances and 11 integer-valued attributes. Among all algorithms, SVM gave the highest accuracy 97.13% with lowest error rate conducted in WEKA data mining tool.

Paper [26] applied 3 machine learning techniques: Naïve Bayes, SVM and RF. Out of them RF is the most appropriate and useful algorithm to give the best accuracy of 99.42% where SVM and NB result 98.8% and 98.24% respectively.

Paper [30] used SVM and SVM ensemble classifier in breast cancer datasets. Further they applied boosting method and RBF kernel based SVM to predict the accuracy for which evaluation parameters are calculated as F-measure, ROC curve etc. on training data to build a model. He found that RBF kernel SVM ensemble based on boosting method performed better accuracy than other classifiers.

In paper [32], a novel hybrid intelligent system on the basis of Association Rule Mining (ARM) and Neural Networks (NN) which utilizes an Evolutionary Algorithm (EA) is introduced to cover the amplitude issue for the finding of breast cancer. ARM enhanced by Grammatical Evolution (GE) is utilized to choose the most useful traits and diminish the admeasurement by extracting sodality among SNPs, while NN is utilized for effective distribution.

In paper [35] authors have designed a new hybrid technique of Grey Wolf Optimizer (GWO) consolidating with decision tree as a classifier for choosing a minimum number of useful genes from the lots of genes to recognize cancer is designed. Implemented technique and other famous classifiers like BPNN, SOM, SVM, C4.S, and PSOC4.S applied to tests on 10 gene expression cancer datasets and outcome demonstrates implemented technique is better than the other present techniques.

In paper [37] selected informative genes from thousands of genes of microarray for classification. The swarm intelligence techniques such as PSO, Cuckoo Search, Shuffled Frog Leaping and Shuffled Frog Leaping with Levy Flight (SFLLF) find the informative genes. K- Nearest Neighbor (k-NN) classifier is used to classify the samples. The best result obtained from k-NN classifier through SFLLF feature selection methods. The data is taken from Kent Ridge Biomedical Data repository. Various datasets like CNS, DLBCL, Lung Cancer, Ovarian Cancer, Prostate outcome, AML/ALL, Colon Tumor etc. are experimented to perform many swarm intelligence techniques viz PSO, CS, SFL, SFLLF to get the best accuracy result. SFLLF outperforms the best result.

Paper [38] performed a unified view of multi-class of SVM for multiple datasets. They explained multi-class loss function, margin function, aggregator operator, fisher consistency of multi-class loss function mathematically. The accuracy result of linear SVM and non-linear SVM are also tabulated.

Papers [39] compare the supervised and unsupervised machine learning classifier for the breast cancer dataset. They identified that supervised ML required to applied on training dataset whereas unsupervised ML does not require to train the model by training dataset. Robust one-class SVM and enhanced one-class SVM are applied for anomaly detection in the dataset and found enhanced one-class SVM is more superior than other.

Paper [40] proposed a cancer site classification framework by investigating somatic mutations through machine learning approaches. They extracted information related to patient information, mutation associated genes, and mutation-associated chromosomes from the database COSMIC (Catalogue of Somatic Mutations in Cancer), also integrated the mutation-associated gene function using gene pathways from the database KEGG (Kyoto Encyclopaedia of Genes and Genomes).

Paper [41] proposed a highbrowed distribution framework to observe typical and anomalous MRI brain images. These days, judgment and analysis of brain tumors depend on syndrome and radiological presence. Magnetic resonance imaging (MRI) is the most critical regulated tool for the gross examination of tumors in the brain. In the present examination, different methods were utilized for the characterization of brain cancer. Under these systems, picture preprocessing, image feature extraction and subsequent classification of brain cancer are effectively performed. At the point when distinctive machine learning strategies: Support Vector Machine(SVM), K-Nearest Neighbor (KNN) and Hybrid Classifier (SVMKNN) is utilized to order 50 pictures, it is seen from the outcomes that the Hybrid classifier SVM-KNN exhibited the most elevated arrangement precision rate of 98% among others.

IV. OBSERVATION

On the basis of survey, we have found that there are so many types of cancers and the treatment of diagnosis and prognosis is different. The behavior and attributes are varying across the cancer types. We have categorized the types of cancer as below.

Ovarian Cancer

The Ovarian cancer is developed only in female. In the female body, there are ovaries that are reproductive glands used to produce eggs for the reproduction. The egg travel from ovaries to the uterus and fertilizes the egg into a fetus. ML techniques had been implemented on ovarian cancer dataset and predicted the accuracy as shown in table 1.

Table 1: Accuracy measured for ovarian cancer using ML Technique.

Type of cancer: Ovarian Cancer			
Ref	**ML Technique**	**Sample**	**Accuracy**
[19], 2014	K-Means with Harmony search	155 (IRIS Data)	97%
[21], 2015	SVM	498	96%
	ANN		93%
	Naïve Bayes		90%

Liver Cancer

The Liver is the largest internal organ in the human body. It lies under the right ribs beneath the right lung. The liver cancer dataset is collected and ML techniques are applied to find the accuracy for the diagnosis of cancer. The accuracy in percentage is shown below in table 2.

Table 2: Accuracy measured for liver cancer using ML technique.

Type of Cancer: Liver Cancer			
Ref	**ML Technique**	**Sample**	**Accuracy**
[9], 2011	Fuzzy Neural Network (FNN)	159	95.35%
[42], 2013	Particle Swarm Optimization	-	93.3%

Colon Cancer

The colon cancer starts in the colon or in the rectum. The first symptom of colon cancer is bloody stools. The common symptoms are seen for colon and rectum both. The ML Techniques are applied to predict the accuracy for the sample of Colon cancer as shown in table 3.

Table 3: Accuracy measured for colon cancer using ML technique.

Type of Cancer: Colon Cancer			
Ref	**ML Technique**	**Sample**	**Accuracy**
[7], 2003	SASOM	62	93.6%
[43], 2008	Fuzzy Granular Support Vector Machine—Recursive Feature Elimination (FGSVM-RFE)	62	99.71%
[22], 2015	KNN using Biogeography- Based Optimization	62	80%

Breast Cancer

The malignant cells often form the tumor and that can easily be seen in the X-Rays in the form of lump. Further the cells start growing into surrounding tissue to distant area of the body.

Table 4: Accuracy measured for breast cancer using MLtechniques.

Type of Cancer: Breast Cancer			
Ref	**ML Technique**	**Sample**	**Accuracy**
[13], 2013	Random Forest	699	99.82%
[23], 2015	DT-SVM	699	91%
[26], 2017	Random Forest	699	99.24%
	SVM		98.8%
	Naïve Bayes		98.24%
[16], 2013	Decision Tree (C4.5)	1189 (Iranian Centre)	93.6%
	ANN		94.7%
	SVM		95.7%
[20], 2014	RepTree (C4.5)	286	71.32%
	Radial Basis Function Network		73.77%
	Simple Logistic		74.47%

The breast cancer datasets are collected from different resources and applied ML techniques in order to predict the accuracy of the cancer diagnosis as shown in table 4. Some of the Machine Learning techniques are performing 100% accuracy to the number of samples and types of features available with the datasets.

Table 5: Various ML Techniques applied on cancer for 100% accuracy

Ref	**ML Technique**	**Types of Cancer**	**Sample**	**Accuracy**
[6], 2008	SVM-RFE	Lymphoma	25	100%
		Breast Cancer	84	
		Colon Cancer	45	
		Lung Cancer	72	
		Ovarian Cancer	39	
[8], 2011	SVM	Liver Cancer	156	100%
[11], 2012	PSO-KNN	Breast Cancer	97	100%
	PSO-SVM			
[18], 2014	SFLLF	Colon	62	100%
		Prostate	136	

The table 5 shows the list of respective classifier and feature selection techniques that are used in several research papers for the prediction of cancer in different aspects. This table also shows that the dataset samples taken to find the accuracy depend upon the instance of data and k-fold cross validation.

The observation is made on the basis of surveying various machine learning algorithms in the prediction of cancer at early stage. Prior to machine learning algorithm, we need to collect the dataset and make them standardized. There may be large data which has no meaning to classify the data uselessly. Sometimes it happens that we are not able to use the entire data then we need to reduce the dimension of dataset attribute.

V. Conclusions

The survey on emerging field of automatic cancer classification and prediction is presented here. It belongs to research area of bioinformatics. From this survey we conclude that, most of the automatic cancer predication systems are based on machine learning concepts including classification and clustering algorithms. This paper presented an extensive review of various ML classification techniques for the prediction of cancer and standard datasets have been used in wide variety of cancer such as ovarian cancer, breast cancer, liver cancer and colon cancer.

A detailed list of results found by many researchers has been tabulated to solve the problems by various computational intelligence techniques. The most successful approach is SVM and combination of SVM technique which gave up to 100% accuracy on a smaller number of training datasets which is not a good prediction in case with large datasets. However, options are available for the possibilities of improvement of predicting the cancer at an early stage. There are many datasets available to explore more for the same. There are large numbers of cancer types available with unknown functions.

ACKNOWLEDGMENT

The authors are thankful to the publishers, researchers for making their resources available and also appreciate the comments and suggestions obtained from the reviewers, which are useful to improve the quality of paper.

REFERENCES

[1] WorldHealthOrganization(WHO),CancerFactSheet,Http://www.Who.int/Mediacentre/Factsheets/Fs297/En/(Accessedon:October/2017),2017.

[2] 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," *in IEEE Transactions on Computational Social Systems*, vol. 3, no. 2, pp. 63–74, June 2016.

[3] Alpaydin E. Introduction to Machine Learning. MIT press; 2009.

[4] Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behav Sci Law. 2019;37(3):214-222.

[5] Marshland, S. (2009) Machine Learning an Algorithmic Perspective. CRC Press, New Zealand, 6-7.

[6] Y. Hu, K. Ashenayi, R. Veltri, G. O'Dowd, G. Miller, R. Hurst and R. Bonner, "A Comparison of Neural Network and Fuzzy c-Means Methods in Bladder Cancer Cell Classification", Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), pp 3461- 3466, ISBN: 0-7803-1901-X DOI: 10.1109/ICNN.1994.374891 http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=37 4891

[7] Sung-Bae Cho, Hong-Hee Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification", First Asia-Pacific Bioinformatics Conference, Adelaide, Australia. Conferences in Research and Practice in Information Technology, Vol. 19. 2003

[8] Rui Xu, Xindi Cai, Donald C. Wunsch II, "Gene Expression Data for DLBCL Cancer Survival Prediction with A Combination of Machine Learning Technologies", Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 2005, pp 894-897, ISBN 0780387406

[9] P. Rajeswari, G. Sophia Reena, "Human Liver Cancer Classification using Microarray Gene Expression Data", International Journal of Computer Applications (0975–8887) Volume 34–No.6, November 2011, pp 25-37.

[10] Jayashree Dev, Sanjit K Dash, Swet Dash, Madhusmita Swain, "A Classification Technique for Microarray Gene Expression Data using PSO-FLANN", International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 09 Sep 2012, ISSN: 0975- 3397, pp 1534-1539.

[11] Barnali Sahu, Debahuti Mishra, "A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data", International Conference on Modeling Optimization and Computing (ICMOC-2012), ELSEVIER Procedia Engineering 38 (2012) pp 27–31.

[12] S. Mishra, C. D. Kaddi, and M. D. Wang, "Pan-cancer analysis for studying cancer stage using protein and gene expression data," in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, 2016, pp. 2440–2443.

[13] Cuong Nguyen, Yong Wang, Ha Nam Nguyen," Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic", **J. Biomedical Science and Engineering, 2013, 6,** pp 551-560, DOI: http://dx.doi.org/10.4236/jbise.2013.65070

[14] Ammu P K, Preeja V," Review on Feature Selection Techniques of DNA Microarray Data", Intl. J. of Computer Applications (0975– 8887) Volume 61–No.12, January 2013, pp 39-44.

[15] B.M.Gayathri, C.P.Sumathi, T.Santhanam, "Breast cancer diagnosis using machine learning algorithms–a survey", International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013, pp 105-112.

[16] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence", Open Access, Journal of Health & Medical Informatics 2013, vol 4, issue 2. ISSN: 2157-7420, http://dx.doi.org/10.4172/2157-7420.1000124

[17] P. Ramachandran, N.Girija, T.Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications (0975–8887), Volume 97– No.13, July 2014, pp 48-53.

[18] Mehdi Pirooznia, Jack Y Yang, Mary Qu Yang, Youping Deng, "A comparative study of different machine learning methods on microarray gene expression data", BMC Genomics, Open Access BioMed Central, 2008, International Conference on Bioinformatics & Computational Biology (BIOCOMP'07) Las Vegas, NV, USA. 25-28 June 2007, DOI: 10.1186/1471-2164-9-S1-S13.

[19] Arunanand T A, Abdul Nazeer K A, Mathew J P, Meeta Pradhant, "A Nature-inspired Hybrid Fuzzy C-means algorithm for Better Clustering of Biological Data Sets", IEEE International Conference on Data Science & Engineering (ICDSE '14), pp 76-82.

[20] Vikas Chaurasia, Saurabh Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.1, January- 2014, pg. 10-22, ISSN **2320–088X**

[21] P. Yasodha, N.R. Anathanarayanan, "Analysing Big Data to Build Knowledge Based System for Early Detection of Ovarian Cancer", **Indian Journal of Science and Technology (IJST),** Vol 8(14), July 2015, ISSN *0974-5645*, DOI: 10.17485/ijst/2015/v8i14/65745

[22] Ammu P K, Siva Kumar K C, Sathish M, "A BBO Based Feature Selection Method for DNA Microarray", International Journal of Research Studies in Biosciences (IJRSB) Volume 3, Issue 1, January 2015, PP 201-204, ISSN 2349-0365

[23] K. Sivakami, "Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model", International Journal of Scientific Engineering and Applied Science (IJSEAS), Volume-1, Issue-5, August 2015, pp 418-429, ISSN: 2395-3470.

[24] Kar, Subhajit, Kaushik Das Sharma, and Madhubanti Maitra, "A particle swarm optimization based gene identification technique for classification of cancer subgroups," in *2nd IEEE International Conference on Control, Instrumentation, Energy and Communication (CIEC)*, 2016.

[25] Hiba Asria, Hajar Mousannifb, Hassan Moatassimec, Thomas Noeld, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", ELSEVIER 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), Procedia Computer Science 83 (2016) pp 1064–1069.

[26] Madeeh Nayer Elgedawy, “Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes”, International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 1 Jan. 2017, Page No. 19884-19889 Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.07

[27] http://www.medicalnewstoday.com/info/cancer-oncology

[28] https://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0072594/

[29] Tanupriya choudhary, Vivek Kumar, Darshika Nigam, Vasudha Vashisht, “ An Innovative Smart Soft Computing Methodology towards Disease (Cancer, Heart Disease, Arthritis) Detection in an Earlier Stage and in a Smarter Way”, International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 3, Issue 4, April 2014, pp 368–388, ISSN 2320-088X.

[30] Min-Wei Huang, Chin-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, Chin-Fong Tsai, "SVM and SVM Ensembles in Breast Cancer Prediction." *PloS one* 12.1 (2017): e0161501. DOI: 10.1371/journal.pone.0161501, January 6, 2017.

[31] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, “Machine Learning Applications in Cancer Prognosis and Prediction”, ELSEVIER Computational and Structural Biotechnology Journal 13 (2015), pp 8-17,DOI: http://dx.doi.org/10.1016/j.csbj.2014.11.005

[32] Boutorh, Aicha and Ahmed Guessoum, “Classication of SNPs for breast cancer diagnosis using neural-network-based association rules,” in *12th International Symposium on Programming and Systems (ISPS)*, IEEE, 2015.

[33] Yu-Xiong Wang, Martial Hebert, “ Learning to Learn: Model Regression Networks for Easy Small Learning”, European Conference on Computer Vision (ECCV 2016) pp 616-634, DOI:10.1007/978-3-319-46466-4_37, Lecture Notes in Computer Science, vol 9910 Springer.

[34] V. Jothi Prakash, L.M.Nithya, “A Survey on Semi-Supervised Learning Techniques”, International Journal of Computer Trends and Technology (IJCTT) vol 8 no.1, Feb 2014. ISSN 2231-2803, pp 25- 29.

[35] M. Vosooghifard and H. Ebrahimpour, “Applying grey wolf optimizer-based decision tree classifier for cancer classification on gene expression data,” in *5th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, 2015, pp. 147–151.

[36] Shikha Agrawal, Jitendra Agrawal, “Neural Network Techniques for Cancer Prediction: A Survey”, 19th Internation Conference on Knowledge Based and Intelligent Inforkation and Engineering Systems, ELSEVIER SciecnDirect Procedia Computer Science 60 (2015) 769-774, DOI: 10.1016/j.procs.2015.08.234

[37] C Gunavathi, K Premalatha, “A Comparative Analysis of Swarm Intelligence Techniques for Feature Selection in Cancer Classification”, Hindawi Publishing Corporation The Scientific World Journal, Volume 2014, Article ID 693831, pp 1-12, ISSN 2356-6140,http://dx.doi.org/10.1155/2014/693831

[38] Urun Dogan, Tobias Glasmachers, Christian Igel, “A Unified View on Multi-class Support Vector Classification”, Journal of Machine Learning Research 17 (2016) 1-32.

[39] Mennatallah Amer, Markus Goldstein, Slim Abdennadher, “Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection”, In Proceedings of the ACM SIGKDD 2013 Workshop on Outlier Detection and Description (pp. 8-15).

[40] Chen, Yukun, et al., “Classification of cancer primary sites using machine learning and somatic mutations,” *BioMed Research International*, 2015.

[41] K. Machhale, H. B. Nandpuru, V. Kapur, and L. Kosta, “MRI brain cancer classification using hybrid classifier (SVM-KNN),” in *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, Pune, 2015, pp. 60–65.

[42] Akanksha Sharma, Parminder Kaur, “Optimized Liver Tumor Detection and Segmentation Using Neural Network”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-5, November 2013, pp 7-10.

[43] Yuchun Tang, Yan-Qing Zhang, Zhen Huang, Xiaohua Hu, Yichuan Zhao, “Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification”, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 12, NO. 6, NOVEMBER 2008, pp 723-730.