Application of Machine Learning for Prediction of Lung Cancer using Omics Data

Arwinder Dhillon, Amrita Kaur, Ashima Singh

Abstract: ***Cancer is one of the deadly diseases across many countries. However, cancer can be cured, if detected at an early stage. Researchers are working on healthcare for early detection and prevention of cancer. Medical data has reached its utmost potential by providing researchers with huge data sets collected from all over the globe. In the present scenario, Machine Learning has been widely used in the area of cancer diagnosis and prognosis. Survival analysis may help in the prediction of the early onset of disease, relapse, re-occurrence of diseases and biomarker identification. Applications of machine learning and data mining methods in medical field are currently the most widespread in cancer detection and survival analysis. In this survey, different ways to detect and predict lung cancer using latest Machine learning algorithms combined with data mining has been analyzed. Comparative study of various machine learning techniques and technologies has been done over different types of data such as clinical data, omics data, image data etc.***

Keywords: Lung Cancer, Omics Data, Images data, Machine Learning, Survival Analysis, Supervised Learning

I. INTRODUCTION

Cancer is a generic term for a group of diseases which may affect multiple or related parts of the body, such as liver, kidney, breast, hair, skin, etc. Cells are the basic unit of a human body which grows and divides and forms new cells when required by the body. However, cancer arises when certain genetic changes disrupt certain natural changes, resulting in uncontrollable development of the cells. Throughout 2018, there were about 17 million new cases of cancer worldwide and it is estimated that by 2030 there will be 22.5 million new cases of cancer every year [1]. Cancer can be modifiable and non-modifiable. Modifiable cancer involves behavioral and wellness habits such as tobacco use, weight gain etc., and non-modifiable genetic factors such as genetic mutations inherited and immune deficiencies. Smoking is the primary risk factor for lung cancer, although studies show that smoking is directly associated with 90% of women and 79% of men [2]. A community of cancerous cells that grow in the lung tissue causes lung cancer. Such rare cells expand hysterically, allowing a malignant tumor to form. Such deviant cells do not behave as normal functioning of a typical lung cell and interfere with normal and healthy lung function. Generally, as this occurs, our body has the right number of each cell as cells usually work by generating indications as to how regularly and in what quantity they have to split. Any kind of mistake in these signals leads cells to behave abnormally and replicate too much, contributing to tumor forming [3]. Each gene is an instruction that tells the cell to make something, usually protein. Genes contain DNA (deoxyribonucleic acid), which contains our unique genetic code and controls cell to act properly. Cell growth and reproduction is controlled by genes for a healthy body. Sometimes a change or disruption may occur in gene called as mutation. Mutation in specific cells causes over production of proteins which forces cell to divide further and so on. Lung cancer is classified as small cell lung cancer and non-small cell lung cancer. In its earliest stages, lung cancer, like other cancers, is most curable before it has spread to other areas of the body. Non-small cell lung cancer is most common and can be categorized into adenocarcinomas of the lungs, squamous cell carcinomas, and large cell carcinomas. SCLC lung cancer is the most violent and fast-paced of all forms. SCLC is strongly associated with smoking cigarettes. The World Health Organization (WHO) carried out research worldwide in 2012, showing that 8.2 million people died in one year from cancer-related disease [4]. The diagnosis and treatment of lung cancer in its early stages may increase the patient's survival rate from 14% to 49%, according to the American lung cancer society [5]. Figure 1 shows the global cancer statistics. This is a driving force in the quest for effective algorithms for better treatment to predict survival rates. Due to the high dimensionality of data, it is a tedious task to extract information from it. Researchers have used numerous algorithms in this area to the best of their ability to use large amounts of data in the medical sciences. The ability of machine learning to use different techniques such as probabilistic, optimization, statistical etc. makes this branch tailor-made for the handling of vast predictive data. With the rapid development of genomic, proteomic and imaging technologies, patient or disease molecular level data can be easily gained.

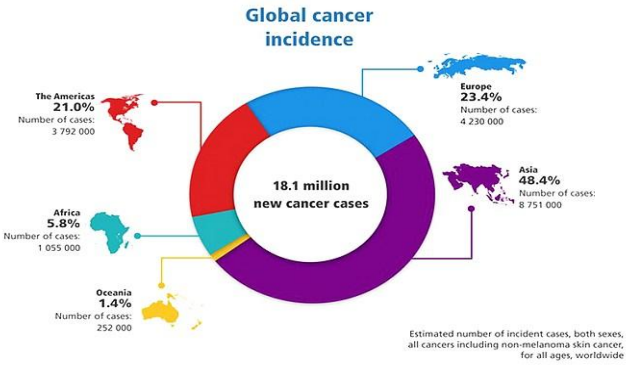

Fig. 1. Global cancer statistics

Revised Manuscript Received on April 1, 2020.

Arwinder Dhillon, Assistant Professor, Computer Science and Engineering Department, at Thapar Institute of Engineering & Technology, Patiala.

Amrita Kaur, research scholar in Computer Science and Engineering Department, at Thapar Institute of Engineering & Technology, Patiala.

Ashima Singh , Leturer in Computer Science and Engineering Department, at Thapar Institute of Engineering & Technology, Patiala

A. MACHINE LEARNING

Utilization of machine learning applications in healthcare is currently essential and mandatory in efforts to renovate the raw data into a beneficial asset. The basic principle of machine learning is to produce certain methods that can accept input data and use different methods of analysis to predict performance while updating outputs as new data become available. Over the years passed, we have been dependent on the macro level information such as environment, patient, tumor etc so the number of variable we have encountered has been less but with today's high dimensional solution to questions along with the technologies in imaging we are besieged with enormous amount of particle level, cell level and scientific variables [6]. With such scenarios human insight and standard information and methods don't normally work. Hence we need to depend on vigorous computational methodologies like Machine Learning.

B. MACHINE LEARNING ALGORITHMS

SUPERVISED LEARNING

Supervised learning is a learning category in which the machine is taught or trained using marked data, which means that data has been previously marked with the correct answer. The machine is supplied with fresh data so that the algorithm evaluates the training data and hence can provide with the correct output from the labelled data [7]. Supervised learning is graded into two classification and regression categories. When the resulting parameter is a category, a classification algorithm is used. If we get a numerical value as the resulting parameter, a regression algorithm is. Naive Bayes, Artificial Neural Networks, Support Vector Machines, Logistic Regression, and Random Forests [8] are examples of supervised learning algorithms. In both the types of supervised learning algorithms, regression and classification, the goal is to find explicit patterns or composition in the data provided to help us generate accurate output data effectively.

UNSUPERVISED LEARNING

Unsupervised learning may be defined as learning in which there is no preceding information about any class of data or any label. In this, machine tries 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 algorithms involved in 12 unsupervised learning are K-mean clustering, Association Rule Mining, Topic Modeling and Dimensionality Reduction Techniques [7].

REINFORCEMENT LEARNING

Reinforcement learning works in such a way that an agent improves its performance by taking feedback from the environment. It repeats this feedback process until it has learned efficiently [7]. The aim is to ascertain the best actions within a particular situation so as to obtain maximum performance. It can lead to more customized and accurate treatments at reduced costs. Reinforcement learning in healthcare is used in areas such as medical image screening for diagnosis detection, medical chat bots, clinical decision making simulation etc.

C. SURVIVAL ANALYSIS

Survival analysis may be defined as a group of algorithms that work to evaluate in resulting variable in the form of time that depicts the period until the occurrence of the event we are interested in. This event can be like death which in machine learning is considered as failure. While regression models cannot work on censored or uncensored data, the survival models can efficiently work on this type of data [9]. Survival models can be either parametric or non-parametric. Some survival models are semi parametric as well. Non parametric models works best when we need a directional view to compare which group has better survival rate while parametric models are used to predict time quantiles using various distributions. Semi parametric works accordingly depending upon the data [10].

The rest of paper is discussed as follows. Section 2 describes the type of data used for lung cancer prediction. The different machine learning algorithms used for lung cancer prediction are described in section 3. In section 4, conclusion is discussed.

II. TYPES OF DATA USED

A. HIGH DIMENSIONAL DATA (OMIC DATA) FOR CANCER PREDICTION

Data is enormously large in various forms and types. There exists high dimensional data, which has proved to be a useful resource in healthcare, because of its molecular level information such as omic data. Omics are a branch of bioinformatics which deals with analysis of molecular profiles. Omic data include Genomics, Transcriptomic, Metabolomics, Epigenetics, Lipid omics, Proteomics, and Glycomics. Our main focus is genomics, which is a field of biology that deals with genome of an individual. Dealing with such data requires efficient procedures and computational power as such volumes of data are not a man's job to cater to. This is where machine learning along with data mining comes into picture. High dimensional data refers to data with extremely high number of dimensions which makes the calculations very difficult. Genetic data is such voluminous data which can be handled by the application of machine learning. We have an outburst of features such that number of feature exceeds the number of observations. In such cases powerful algorithms are necessary for efficient utilization of such resourceful data. The following authors work on omics and clinical data. Dokyoon Kim et.al [11] presented a graph-based approach for integration of multi-omics data and genomic data for the prediction of clinical cancer outcomes. Clinical and genomic data profiles for lung cancer dataset from TCGA data portal were taken and experiment was performed. The result proved that the proposed approach performed best with Area under Curve value of 0.7866. Denis Bertrand et.al [12] proposed integration framework OncoIMPACT to eliminate the patient specific genes on 1000 samples of genomic profiles. Experiment was performed and results showed that proposed approach worked well by accurately eliminating the patient specific genes.

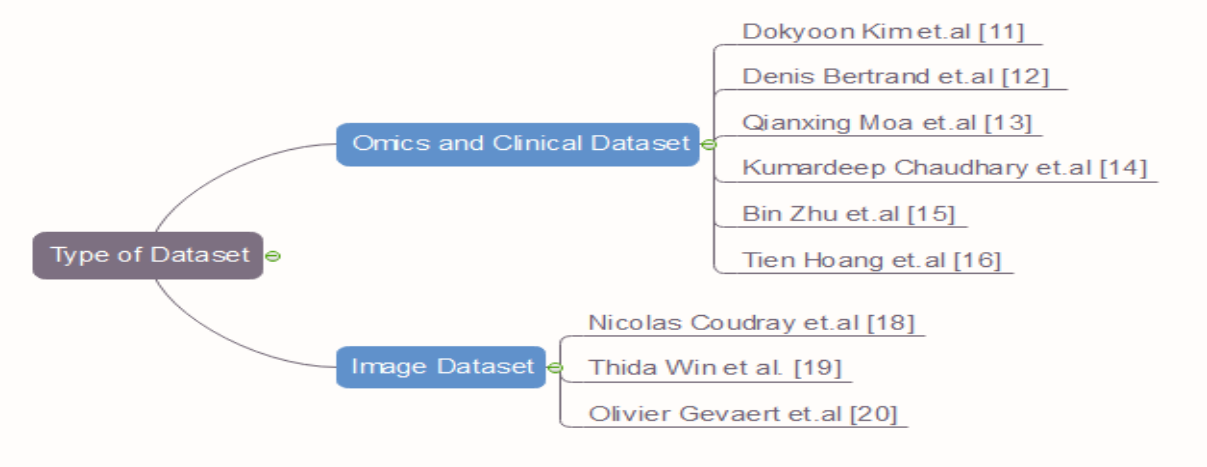

Fig. 2. Types of data used for lung cancer

Qianxing Moa et.al [13] proposed a framework of continuous and discrete joint modelling for pattern discovery and gene-identification. A sample of 2000 tumor patients has been chosen and experiment was performed. Results showed that the proposed approach performed with best accuracy. Kumardeep Chaudhary et.al [14] used deep learning framework based on hepatocellular carcinoma (HCC) to divide survival patients in six cohorts. A sample of 360 HCC patients were taken from cancer genome atlas dataset. Experiment was performed and results show that the Deep learning framework accurately divide two subgroups with p-value of 7.3148*e-6 and concordance index value of 0.68. Bin Zhu et.al [15] used multi-kernel omics machine learning method for prediction of lung cancer on 3328 patients. The experiment was conducted and the results showed that the proposed framework worked well with a 0.86 c-index value. Tien Hoang et.al [16] identified clinical factors for predicting survival in chemotherapy-naive patients with advanced non-small cell lung cancer (NSCLC) treated with chemotherapy regimens of the third generation. Patient samples were taken and analysis of cox regression was performed. It is proved from the result that the proposed approach performed well with hazard ratio value of 1.26. 243 patients are passed to PET/CT test and results proved that the proposed approach works well with hazard ratio of 0.54. The different types of data for lung cancer prediction is shown in the table 1 and figure 2.

Table-I: Different types of data used

AUTHOR	Type of Data		
	Clinical Data	OMIC DATA	IMAGE DATA
Dokyoon Kim et.al [11]	✓	✓	_
Denis Bertrand et.al [12]	_	✓	_
Qianxing Moa et.al [13]	_	✓	_
Kumardeep Chaudhary et.al [14]	✓	✓	_
M. Teverovskiy *et al [17]*	_	_	✓
Bin Zhu et.al [15]	✓	✓	_
Nicolas Coudray et.al [18]	_	_	✓
Thida Win et al. [19]	✓	_	✓
Olivier Gevaert et.al [20]	_	✓	✓
Tien Hoang et.al [16]	✓	_	_

III. MODELS USED

A. Bayesian networks

A Bayesian network is an illustration of a joint distribution of probability of a set of random variables with a potential fundamental relationship. The network entails nodes that denote random variables, edges between pairs of nodes that represent the principal relationship of these nodes, and a conditional distribution of probability in each node. [21,22]. Bayesian network is computed by using the chain rule which is given by:

$$P\ (x_1 \ldots . x_n) = \prod_{1}^{n} P(x_i | Parents(x_i)) \qquad (1)$$

A Bayesian network is a graph in which includes:

- A random variable set which acts as nodes for the graph.
- A directed links set A to B to connect the nodes where A implies that it has a direct impact on B.
- A conditional probability table (CPT) which determines the impact of parents on that node.
- A directed acyclic graph with no directed cycles.

B. Support vector machines

A Support Vector Machine (SVM) classifies by finding the hyper plane that maximizes the margin between the two classes. Support vectors are the vectors (cases) that define the hyper plane. It is a fast and reliable classification algorithm with a limited amount of data that performs very well.

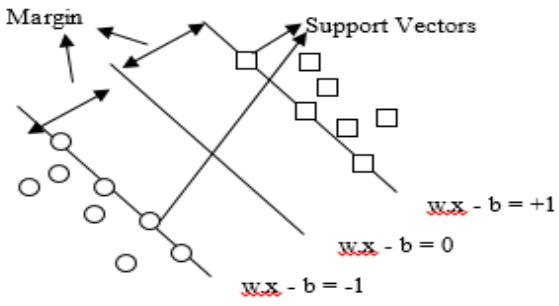

Fig. 3. Structure of SVM

The figure represents the structure of SVM in which a line is used to distinguish the different
classes into two parts [23]. It consists of three lines where line w.x - b = 0 represents the margin line because it is used to separate two patterns. The lines w.x-b = 1 and w.x-b = -1 are the rows on both sides of the margin row. Together, these three lines create the hyper plane separating the given patterns and these separated patterns are known as support vectors.

C. Artificial neural networks

The artificial neural network is a computational system that is biologically influenced and patterned after the human brain's neuron network [24]. It is also possible to consider artificial neural networks as learning algorithms that model the relationship between input and output. An artificial neural network transforms input data through the application of a nonlinear function to a weighted input sum. The transformation is called a neural layer and the function is called a neural unit. First layer's intermediate outputs, called features, are used in the next layer as the input. The neural network learns multiple layers of nonlinear features (such as shapes and edges) through repeated transformations, which it then combines to construct a prediction in a final layer (of more complex objects). The neural network learns by adjusting a network's weights or parameters to reduce the discrepancy between the neural network's predictions and the desired values. The structure of neural network is shown in figure 4.

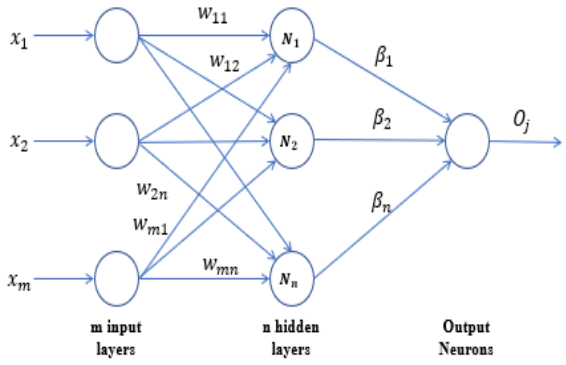

Fig. 4. Structure of Neural Network

D. Decision tree

Decision Tree are forms of supervised machine learning where the information is constantly divided according to a pattern that fits. Two entities describe the tree, namely nodes and leaves of the decision. The decision nodes are the nodes where the data is to be split into child nodes and leaves [26]. The structure of decision tree is shown below.
In order to check which node will go to be used as the left node, it is important to determine the information gain value, i.e. the highest information gain value node is chosen. To calculated gain, entropy is needed given as:

$$H\,(S)\sum_{c\varepsilon C} -p(c)log_2\,p(c) \qquad (1)$$

Here H(S) is the entropy, C is the set of classes, and S is the data, p (c) is the probability of C with respect to S.
Use this entropy to calculate Information gain which is given below:

$$IG(A,S) = H(S) - \sum_{t\varepsilon T} p(t)H(t) \qquad (2)$$

Here, H(S) defines entropy, T is the subset on which decision to be made, p(t) gives the probability of T with respect to S, H(t) is entropy on subset T.

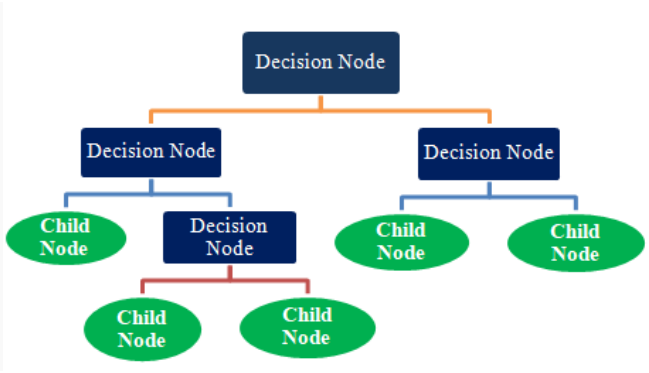

Fig. 5. Structure of decision tree.

E. Random Forest

Random Forest is an algorithm for supervised learning. Random forest builds and merges several decision trees to obtain a more reliable and precise prediction [25]. In regression, the output is averaged for each tree to get the output for input and in classification, voting is done by each tree and output is decided according to maximum votes obtained by a class. It is considered as an ensemble of various simple decision trees. In this, complete dataset is taken which is partitioned in bootstrap samples that we bag them in k groups randomly. On each sample, decision tree algorithm runs and based on the results, voting is performed. Based on the voting, the one with highest vote is chosen as a classification result. The structure of random forest is shown in figure 6.

F. Superpc

Supervised principal components is a simplification of Principal Components (PC) Regression. The initial PC are the regular amalgamation of the structures that capture the course of principal variation in a dataset. To find linear combinations that are associated to an outcome variable, we calculate univariate scores for each gene and then keep hold of only those features whose score surpass a threshold.

G. Survival models

Survival models can be either parametric or non-parametric. Some survival models are semi parametric as well. Non parametric models work best when we need a directional view to compare which group has better survival rate while parametric models are used to predict time quantiles using various distributions. Semi parametric works accordingly depending upon the data. The work on machine learning models for lung cancer prediction is done by following authors. Andre Dekker et.al [22] proposed an approach for "Survival Prediction in Lung Cancer Treated with Radiotherapy" using Bayesian Networks. Author contrasted use of Bayesian networks (BN) with support vector machines (SVM) on a dataset of 322 lung cancer patients.

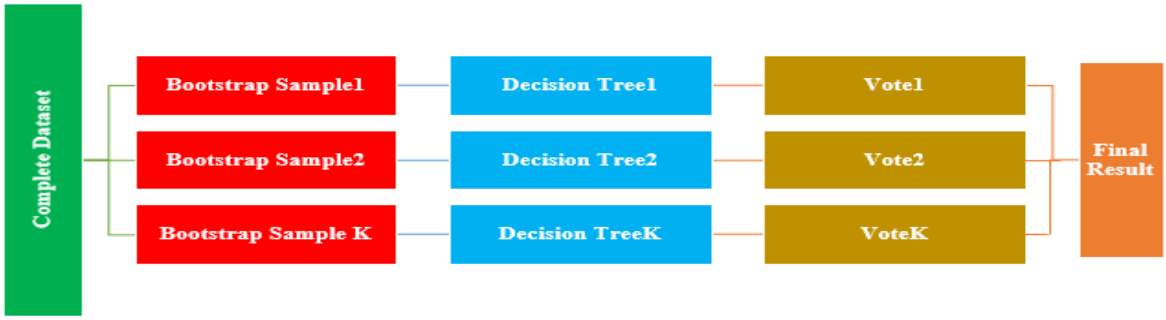

Fig. 6. Structure of decision tree

The performance parameters suggest that the BN outperforms with 80% accuracy. Hege M Bøvelstad et.al [27] proposed a technique for Survival prediction using clinico-genomic models. Author integrate clinical covariates with genomic data and train the integrated data with Cox regression model. It is evident from the results that integrated data gave us considerably better accuracy as compared to other genomic data alone. Prediction models can be further improved by refining the data set and hence the better results. Eric Bair et.al [28] discussed Semi-Supervised methods for the prediction of survival of patients from gene expression data. This method operates fine irrespective of the number of sub classes of tumor. The author discussed the use of centroid method in case of existence of subtype of cancer exists. Trevor Hastie et.al [29] discussed prediction of lung cancer by supervised principal components. This technique is especially useful when amount of variables used is large. The Cox model used in this work needs a further setting and improvement. The drawback pointed out in this paper is that the latent variable used may not be feasible in actual practice. Shraddha Deshmukh [30] presented another approach called Pruned Fuzzy Min-Max (FMM) Neural Network for the diagnosis of Lung Cancer. The only difference with pruned fuzzy min max neural network and Pruned FMM is the quantity of hyper boxes. Pruned FMMNN has less number of hyper boxes which gives in turn helps in diminishing the number of rules for classification analysis. V.Krishnaiah et.al [31] discussed data Mining techniques comprising Bayesian classifiers and Naive Bayesian, Artificial Neural Networks (ANN) and Decision Tree (DT) for the detection of Lung Cancer. The models were train and experimental results prove that the BN performed best with 85% accuracy as compared to ANN and DT. Despite these algorithm's success Lung cancer predictions can be further improved using Clustering and Association Rule Mining techniques. Konstantina Kourou et.al [32] proposed an approach for cancer prognosis and prediction using Machine learning algorithms such as BN, NN and SVM. From the results the author infers that various techniques applied for feature selection can noticeably affect the result when applied to diverse data with high throughput. Joseph A. Cruz et.al [33] discussed various types of methods used for cancer prediction along with multi-dimensional data that can be used to predict cancer and their performance parameters in 18 comprehensive ways. Enhancement in experimental setup and procedure followed would surely augment the overall quality and accuracy of classifiers. Xiao Zang et.al [34] proposed computational advances to examine the morphological features of digital images for NSCLC patients. A sample of Pathological images from 523 ADC patients and 21,511 SCC were analyzed and extracted features were used to predict survival outcomes in ADC and SCC patients. Xueyan Mei [35] used a random forest and reliefF algorithm to predict survival in non-small cell lung cancer patients. To predict 5-year survival in patients with NSCLC, feature selection algorithms along with various machine learning tools were used. The findings show that the suggested model is more reliable than other methods of selecting features. Jung Hun Oh et.al [36] addressed the application of Radiation Pneumonitis in patients with lung cancer. Several standard classification algorithms were used according to their threat to identify different groups. The performance parameter used was MCC in which Kernel SVMs displayed higher value than linear SVM. Yang Xie et.al [37] suggested a background correction approach by using current RMA context correction to include additional information such as negative control beads. Author has considered approaches to estimate parameters such as non-parametric, maximum likelihood estimation, and Bayesian from which the methods of Maximum likelihood and Bayes tend to yield the best results. The experiment was conducted on the data set of the illumine bead array. The use of deep learning algorithms to diagnose lung cancer was discussed by Wenqing Sun et.al [38]. The author tested the feasibility of using deep learning algorithms in this field through the implementation of three deep learning algorithms: Convolutional Neural Network (CNN), Deep Belief Networks (DBN), Stacked De-Noising Auto-encoder (SDAE) with a maximum accuracy of 81%. Stephan Wienert [39] suggested a Virtual Microscopy Images method for the identification of cell nucleus. Author described a minimal model approach using negligible 19 of previous information. This technique can also detect contours without taking their form into account. The result was a 50.908 accuracy value and a 50.859 recall value. Comparative analysis on cancer prediction research using different machine learning algorithms was performed along with their findings in Table 2.

Table-II: Comparative study of related work

Author	Method /algorithm	Result
Andre Dekker et.al [22]	Bayesian networks (BN), Support vector machines (SVM).	BN model outperforms the SVM with accuracy of 85%.
Hege M Bøvelstad et.al [27]	Cox Models	Integrated clinical dataset with genomic data performed better results as compared to genomic data alone.
Eric Bair et.al [28]	Semi-Supervised Methods	Supervised principal component Analysis produces best result with mean square error value of 0.06.
Shraddha et.al [30]	Fuzzy Min- Max Neural Network.	Pruned FMM gives an accuracy of 94% when 8 hyper-boxes were used.
V.Krishnaiah et.al [31]	BN, DT, ANN	BN model gives an accuracy of 81% which is 5% more than the accuracy obtained by ANN and DT
Kourou et.al [32]	ANN, BN, SVM, and DT	ANN performed best with an accuracy of 83%.
Joseph A. Cruz et.al [33]	Machine Learning Algorithms	Identified a variety of developments in the types of machine learning approaches used and their overall success in predicting cancer susceptibility or outcomes.
Xiao Zang et.al [34]	Statistical models	In the test set, the ADC and SCC patients are categorized into high and low risk groups based on the estimation of morphological characteristics extracted.
Xueyan [35]	DT and RF.	Random forest works well with an accuracy of 80%.
Jung Hun Oh et.al [36]	Classification algorithms in the machine learning.	Kernel SVMs gives 0.89 MCC values which is 5% higher than linear SVM.
Yang Xie et.al [37]	Maximum Likelihood Estimation (MLE) and Bayesian estimation, non-parametric estimator.	Maximum likelihood and Bayes methods performed better results as compared to non-parametric method.
Wenqing et.al [38]	CNN, DBNs, SDAE	CNN model gives an accuracy of 80% which on compared with DBN and SDAE shows an improvement of 2% respectively
Stephan Wienert et.al [39]	‘‘minimum-model’’ cell detection and segmentation technique to detect tumor.	The proposed approach achieves an precision and recall value of 0.908 and 0.859 respectively.

IV. CONCLUSION

In this survey we compare numerous broadly used machine learning classification algorithms. It is expected that machine learning systems can increase the speed and accuracy of diagnosis and clinical decision making among doctors, thereby decreasing costs, saving time and refining the health of patients. Data sources with large dimensions such as genomic data has overshadowed other sources of data. The paper reviews the major algorithms related to prior work done in the area of cancer prediction considering the different types of data taken along with the method used and their respective limitations. Comparative study of how some algorithms work better for certain purpose has been done. The main focus is how when we consider different or heterogeneous data results have been superior. The insight and comparisons of the recent research done in this field provides useful information which can be used to predict cancer in their early stage.

REFERENCES

1. A.M. Cryer, A.J. Thorley, “Nanotechnology in the diagnosis and treatment of lung cancer”, in Pharmacology and Therapeutics, 2019.
2. “What are the risk factors of lung Cancer”, [online]. Available: https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm, last accessed 2019/08/09.
3. “What is Cancer?”, [online]. Available: https://www.cancercenter.com/what-is-cancer, last accessed 2019/06/17.
4. “Cancer”, [online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer, last accessed 2019/07/2.
5. “Lung Cancer fact Sheet”, [online]. Available: https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html, last accessed 2019/07/06.
6. A. Dhillon, A. Singh, “Machine Learning in Healthcare Data Analysis: A Survey”, in Journal of Biology and Today’s World, 2019, 8(2), pp. 1-10.
7. K.P. Linthicum, K.M. Schafer, J.D. Ribeiro, “Machine learning in suicide science: Applications and ethics”, in Behav Sci Law, 2019, 37(3), pp. 214-222.
8. P. Kaur, N. Sharma, A. Singh, and B. Gill, “CI-DPF: A Cloud IoT based Framework for Diabetes Prediction”, In IEEE, 2018, pp. 654-660. [Dig. 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)]
9. Jr. Miller, R.G., “Survival analysis”, John Wiley & Sons, 66, 2011.
10. E.T. Lee and J. Wang, “Statistical methods for survival data analysis”, John Wiley & Sons, 476.
11. D. Kim, J.G. Joung, K.A. Sohn, H. Shin, Y.R. Park, M.D., Kim J.H Ritchie, “Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction” in Journal of the American Medical Informatics Association, 2014, 22(1), pp.109-20.
12. D. Bertrand, K.R. Chng, F.G. Sherbaf, A. Kiesel, B.K. Chia, Y.Y. Sia, S.K. Huang, D.S. Hoon, E.T. Liu, A. Hillmer and N. Nagarajan, “Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles” in Nucleic acids research, 2015, 43(7), pp. e44-e44.
13. Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C., Shen, R. Sander, “Pattern discovery and cancer gene identification in integrated cancer genomic data” in Proceedings of the National Academy of Sciences, 2013, 110(11), pp. 4245-4250.
14. K. Chaudhary, O.B. Poirion, L. Lu and L.X. Garmire, “Deep learning–based multi-omics integration robustly predicts survival in liver cancer” in Clinical Cancer Research, 2018, 24(6), pp.1248-1259.
15. B. Zhu, N. Song, R. Shen, A. Arora, M.J. Machiela, L. Song, M.T. Landi, D. Ghosh, N. Chatterjee, V. Baladandayuthapani and H. Zhao, “Integrating clinical and multiple omics data for prognostic assessment across human cancers”, in Scientific reports, 2007, 7(1), p.16954.

16. T. Hoang, R. Xu, J.H. Schiller, P. Bonomi and Johnson D.H., "Clinical model to predict survival in chemonaive patients with advanced non–small-cell lung cancer treated with third-generation chemotherapy regimens based on Eastern Cooperative Oncology Group data", in Journal of Clinical Oncology, 2005, 23(1), pp.175-183.
17. M. Teverovskiy, V. Kumar, J. Ma, A. Kotsianti, D. Verbel, A. Tabesh, H.Y. Pang, Y. Vengrenyuk, S. Fogarasi and O. Saidi, "Improved prediction of prostate cancer recurrence based on an automated tissue image analysis system", In 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821), 2004, pp. 257-260.
18. N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A.L. Moreira, N. Razavian and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning", in Nature medicine, 2018, 24(10), p.1559.
19. T. Win, K.A. Miles, S.M. Janes, B. Ganeshan, M. Shastry, R. Endozo, M. Meagher, R.I. Shortman, S. Wan, I. Kayani and P. Ell, "Tumor heterogeneity as measured on the CT component of PET/CT predicts survival in patients with potentially curable non-small cell lung cancer", in Clinical Cancer Research, 2013.
20. O. Gevaert, J. Xu, C.D. Hoang, A.N. Leung, Y. Xu, A. Quon, D.L. Rubin, S. Napel and S.K. Plevritis, "Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results", in Radiology, 2012, 264(2), pp.387-396.
21. H. Guo and W. Hsu, "A survey of algorithms for real-time Bayesian network inference", In Join Workshop on Real Time Decision Support and Diagnosis Systems, 2002.
22. A. Dekker, C. Dehing-Oberije, Ruysscher De D., P. Lambin, K. Komati, G. Fung, S. Yu, A. Hope, W. De Neve and Y. Lievens, "Survival prediction in lung cancer treated with radiotherapy: Bayesian networks vs. support vector machines in handling missing data", In 2009 International Conference on Machine Learning and Applications, 2009 (pp. 494-497). IEEE.
23. A. Pradhan, "Support vector machine-A survey. International Journal of Emerging Technology and Advanced Engineering", 2012, 2(8), pp.82-85.
24. S. Haykin, "Neural Networks and Learning Machines", in 3/E. Pearson Education India, 2010.
25. G. Manogaran, D. Lopez, "A survey of big data architectures and machine learning algorithms in healthcare", in International Journal of Biomedical Engineering and Technology, 2017, 25(2-4), 182-211.
26. S.R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology", IEEE transactions on systems, man, and cybernetics, 1991, 21(3), pp.660-674.
27. M.B. Hege, N. Stale, and O. Borgan, "Survival prediction from clinico-genomic models-a comparative study", in BMC bioinformatics, 2009, 10 (1), pp. 413.
28. B. Eric and T. Robert, "Semi-supervised methods to predict patient survival from gene expression data", in PLoS Biol, 2004, 2(4), pp. E108.
29. B. Eric, H. Trevor, P. Debashis, and T. Robert, "Prediction by supervised principal components", in Journal of the American Statistical Association, 2006, 101 (473).
30. D. Shraddha, "Diagnosis of Lung Cancer Using Pruned Fuzzy Min-Max Neural Network", 2016.
31. V. Krishnaiah, G. Narsimha, N. Subhash, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, 2013, 4 (1), 39 – 45.
32. K. Kourou, T.P. Exarchos, K.P. Exarchos, Karamouzis M.V. and Fotiadis, D.I., "Machine learning applications in cancer prognosis and prediction", in Computational and structural biotechnology journal, 2015, 13, pp.8-17.
33. A. Joseph, S. David, "Applications of Machine Learning in Cancer Prediction and Prognosis", 2 Issue published, 2006.
34. X. Luo, X. Zang, L. Yang, J. Huang, F. Liang, Rodriguez, "Comprehensive computational pathological image analysis predicts lung cancer prognosis", in Journal of Thoracic Oncology, 2017, 12(3), pp.501-509.
35. M. Xueyan, "Predicting Five-year Overall Survival in Patients with Non-small Cell Lung Cancer by ReliefF Algorithm and Random Forests", IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (2017)
36. J.H. Oh, R. Al-Lozi and I. El Naqa, "Application of machine learning techniques for prediction of radiation pneumonitis in lung cancer patients", In 2009 International Conference on Machine Learning and Applications (pp. 478-483). IEEE (2019).
37. Y. Xie, X. Wang and M. Story, "Statistical methods of background correction for Illumina BeadArray data", in Bioinformatics, 2009, 25(6), pp.751-757.
38. W. Sun, B. Zheng and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms", In Medical imaging 2016: computer-aided diagnosis (Vol. 9785, p. 97850Z). International Society for Optics and Photonics, 2016.
39. S. Wienert, D. Heim, K. Saeger, "Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach", in Scientific reports, 2012, 2.

AUTHORS PROFILE

Arwinder Dhillon is a research scholar in Computer Science and Engineering Department, at Thapar Institute of Engineering & Technology, Patiala. Her research interest includes machine learning and data analytics, computational bioinformatics, cloud computing and IoT applications.

Amrita Kaur is Leturer in Computer Science and Engineering Department, at Thapar Institute of Engineering & Technology, Patiala. She is pursuing Ph.D. in Computer Engineering and holds a Master's Degree, with distinction in Computer Science. She has nine research publications in reputed peer reviewed journals and conferences. Her research interest includes machine learning and data analytics, image processing, medical image processing.

Ashima Singh is Assistant Professor in Computer Science and Engineering Department, at Thapar Institute of Engineering & Technology, Patiala. Ashima Singh was appointed, in 2006, to Computer Science and Engineering Department as Assistant Professor. She has earned Ph.D. in Computer Engineering and holds a Master Degree, with distinction in Computer Science. She has more than 12 years of experience in teaching and research both at UG/PG levels. She has guided 35 theses in M.E. Software Engineering and M.E. Computer Science and Engineering. She has more than 45 research publications in reputed peer reviewed journals and conferences. She is professional member IEEE and Branch Counselor, IEEE Student Chapter at Thapar Institute of Engineering & Technology. Her research interest includes machine learning and data analytics, computational bioinformatics cloud computing and IoT applications.