ML Assistance in Cancer Detection & Treatment

Shivani Gupta \*Megha Gupta Nisha Garg

*Department of Computer Science Department of Computer Science Department of Computer Science*

*MSCW, University of Delhi, India MSCW, University of Delhi, India MSCW, University of Delhi, India*

Uttar Pradesh, India Delhi, India Delhi, India

[guptashivani312002@gmail.com](mailto:guptashivani312002@gmail.com) [meghabis@gmail.com](mailto:meghabis@gmail.com) [garg98564@gmail.com](mailto:garg98564@gmail.com)

***Abstract:* - Cancer is one of the most dreaded diseases of human beings and is a major cause of death all over the globe. More than a million Indians suffer from cancer and a large number of them die from it annually. ML is widely used in the early diagnosis and prognosis of cancer. A variety of machine learning algorithms, including Artificial Neural Network, Bayesian Networks, Support Vector Machines and Decision Tress have been widely used in cancer research for the development of predictive models which are trained by the researchers to give effective and accurate decision. Machine Learning is widely used in the treatment of cancer. Machine Learning is an excellent tool for finding relationships between variables in your data that are too complex for human to economist. This work presents an insight on how machine learning technology is contributing in the field of healthcare especially in cancer diagnosis and its treatment.**

***Keywords: - Cancer detection, Machine Learning, Bayesian Networks, Decision Trees***

I. INTRODUCTION

World health organization stated that cancer is a main source of death around the world. In the year 2020, it was the main reason for almost 10 million deaths. Accurately predicting a tumour remain a challenging task for many physicians. The emergence of new medical technologies and enormous amount of patient data have motivated the path for the development of new strategies in the prediction and detection of cancer.

Normal cells show a property called contact inhibition by virtue of which contact with other cells inhibits their uncontrolled growth. A Cancer cell appears to have lost this property. As a result of this, cancerous cells just continue to divide giving rise to masses of cells called tumors. Tumors are of two types: benign and malignant. Benign tumors normally remain confined to their original location and do not spread to other parts of the body and cause little damage. In malignant tumors, cells grow very rapidly, invading and damaging the surrounding normal tissues. As these cells actively divide and grow, they also starve the normal cells by competing for vital nutrients.

Machine learning is one way scientist are predicting cancer before it ever happens by analyzing patterns, interaction of genes and other factors and the relationship to cancer development. Machine learning is one of the common types of artificial intelligence (AI). it tries to find the pattern from the user given data and process it. The vast data from the browsers in today’s world helps machine to learn with a great accuracy, which results in the increasing growth of accurate output.

Machine learning is of three types: Supervised machine learning, Unsupervised machine learning and Reinforcement machine learning.

Before using any algorithm, it requires a dataset that contains lots of information to analyze patterns and images for the treatment of cancer. These predictive models are trained by using these datasets. In supervised machine learning, machine learns under guidance. In this learning, the training data acts as a teacher to the machine, which is provided with a labeled data. *Regression* and *Classification* problems are solved by supervised learning. In unsupervised learning, the machine had to learn without anyone’s supervision or without anyone’s direction and in this learning, the machine does not feed with the training data. Machine has to figure out the hidden part by comparing the previous given data, no mentor is used to solve the problems as the algorithm in unsupervised learning is only fed by labelled input and it has to get a way to find the output by discovering trends and association. *Association* and *Clustering t*ypes problem are solved by unsupervised learning technique. In reinforcement machine learning, machine learn by hit and trial method and transit from one state to another. Whole reinforcement learning is testing and learning phase. Input depends on the action we take so the other generated output depends whether it receives a punishment or a reward as it follows the try and error method.

Now a days machine learning is used in healthcare domain for number of ways as shown in Fig1. ***TREWS*** [1] is one such algorithm of ML. It stands for targeted real time early warning system. This algorithm is being used at Howard County General Hospital and Suburban Hospital. It is integrated into their electronic medical records system of The Johns Hopkins Hospital, Johns Hopkins Bayview Medical Center and Sibley Memorial Hospital in 2019. It given doctors a second pair of reliable eyes.

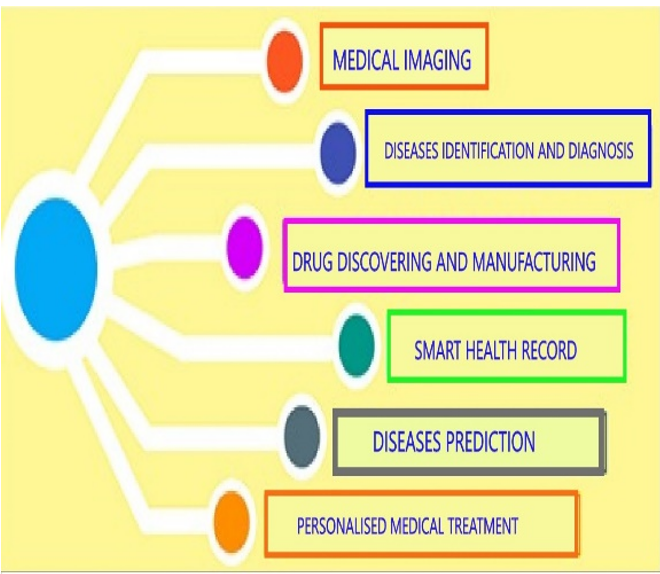


Fig 1: ML Applications in Healthcare domain

II. LITERATURE SURVEY

With the advent of new technologies in the field of medicine, large amount of cancer data has been collected and are available to the medical research community. Recently in the Nov 2020 [2], Nvidia one of the leading USA based multinational technology company announced the intension to build an AI super computer for medical research and drug delivery.

In [3] author recommended use of machine learning, classification and data mining as an effective way to classify the data especially when in biomedical based data, a classification decision is to be made. Millions of deaths has been recorded due to this dreadful disease which can be prevented by usage of new technologies.

According to [4], various machine learning algorithms were applied for the tumour stage prediction of the colon cancer by considering the Tumour Aggression Score (TAS) as a prognostic factor. Performance of different machine learning algorithms were evaluated using fivefold cross validation.

In work [5], author examined, reviewed, categorized, and responded to current breakthroughs in human body cancer detection utilising machine learning approaches for breast, brain, lung, liver, skin cancer, and leukaemia. The paper also shows how machine learning with supervised, unsupervised, and deep learning techniques can help with cancer diagnosis and cure.

III. MACHINE LEARNING ASSISTANCE TO CANCER

This section discusses various machine learning approaches that are nowadays are in use of cancer detection and treatment.

1. *Decision Tree*

Decision Tree is a type of classification algorithm which comes under supervised learning technique. Decision tree is a graphical representation of all possible solutions to a decision based on certain conditions. This technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy [2], [6], [7].

*Decision Tree in diagnosis & prognosis of cancer:*Decision tree has been applied in medical for many purposes like diagnosis of various chronic disease, Predicting Risk of Mortality, feature selection to improve classification accuracy, for reduction of the diagnosis cost. Using supervised learning, it is now simple to classify a tumour as benign or malignant as shown in Figure 2.

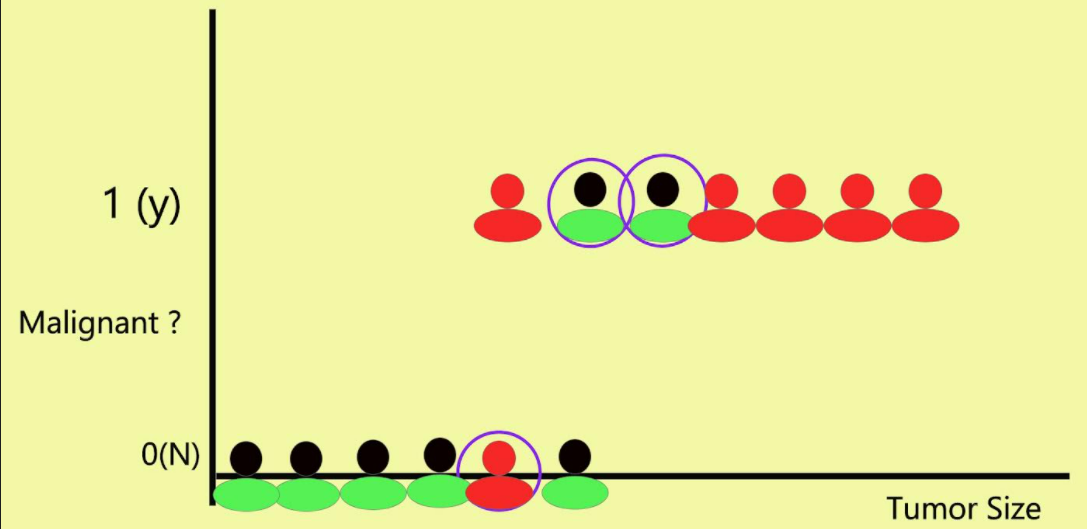


Fig 2: A Tumor classification as malignant or not.

1. *ARTIFICIAL NEURAL NETWORK (ANN)*

ANN was inspired by biological neuron networks present in our brain. Before using any neural network, it requires a dataset that contains lots of information to analyse. In this case, the training data set contains information. This is the data that is needed to train the neural network. A neural network is composed of an input layer, one or more hidden layers and output layers. The neural network begins to learn from the data by adding input data at the input layer. It then flows through the hidden layers eventually arriving at the output layer [2], [6], [7].

*ANN in diagnosis & prognosis of cancer:*Artificial neural networks have featured in a wide range of medical journals, often with promising results. The software development exploits the potential of human intelligence such as reasoning, making decision, learning by experiencing and many others areas

1. *BAYESIAN NETWORK LEARNING*

Bayesian Networks Learning is a graphical model that is used to compute uncertain domains by using the concept of probability that represents conditional dependencies between random variables. Bayesian Networks Learning shows estimation of probability rather than prediction. It is represented by Directed Acyclic Graph (DAC) consisting of nodes and edges [2], [6], [7].

*BN in diagnosis & prognosis of cancer:*Bayesian networks are used in Predicting the prognosis of breast cancer by integrating the clinical and microarray data. The main advantage of this probabilistic model is that it allows integrating these data sources in several ways and that it allows investigating and understanding the model structure and parameter

1. *SUPPORT VECTOR MACHINES (SVM)*

Support Vector Machines are today’s most powerful classification algorithm in terms of predictive accuracy. It is useful in solving both classification and regression problems. The SVM algorithm as a relatively new classification or prediction method has been developed as a result of the collaboration between the statistical and the machine learning research community. It looks at the extreme of the data sets and draws a decision boundary also known as a hyper plane [2], [6], [7].

*SVM in diagnosis & prognosis of cancer:*The efficiency of the Support Vector machine recognition obtained is nearly 97%. This high rate of accuracy can be utilized to support the Doctor’s decision to avoid Biopsy

The main objective of ML techniques is to produce a model which can be used to perform classification, prediction, estimation. It provides a better model to predict and diagnosis the diseases. It can allow us to examine through imaging and gene expression. In past, physicians used populations-based data include family, history, age, weight, diet, high risk habits. But this data is very small and does not provide enough information to made decisions.

With the advancement of High Throughput Technologies, huge amount of data is available and with the help of imaging technologies which proves themselves in healthcare, identification of pattern from complex data set and produce robust results. The features used in data are age, tumour size, tumour location, thickness, TNM staging, haemoglobin, WBC etc [3]. On the basis of TNM staging, ML techniques performed better. There are many papers published which shows applications of ML techniques in healthcare. This is increasing day by day. There is a growing trend of using SVM and BN due to better prediction. In [8] the data set used has information about DNA methylation, gene expression and exon expression. By examining gene, we can predict about cancer like which gene is responsible etc. In [7], Breast Cancer Coimbra dataset from ML Repository was used. The features in the dataset are age, BMI, glucose, insulin, leptin, HOMA etc. These features considered from 52 people which are healthy and 64 women having breast cancers. In [2], Wisconsin Breast Cancer Dataset is used from ML Repository. In this dataset 241 cases are malignant and 458 are benign. K cross fold Validation method is used for testing of predictive models. On comparing, BN is better than other in terms of recall and precision and RF is better in terms of ROC area.

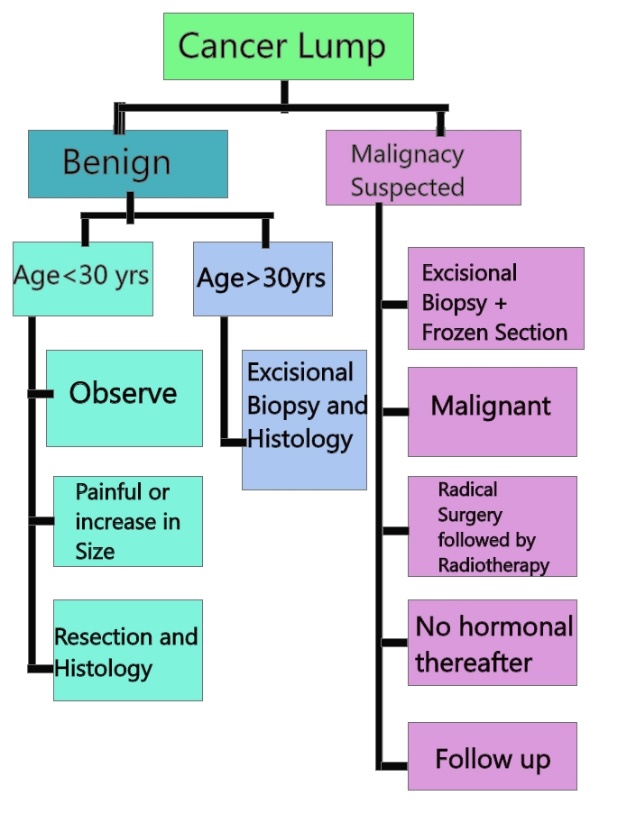


Fig 3: Cancer Diagnosis

Prognosis/prediction of cancer is concerned with three predictive tasks: (i) prediction of cancer susceptibility (ii) prediction of cancer survivability (iii) prediction of cancer recurrence. Figure 3 shows how a lump could be diagnosed as cancer lump and actions taken thereafter. Susceptibility deals with chances of developing a tumour, recurrence deals with chances of regeneration of tumour after treatment and survivability deals with chances of survival of patients after diagnosis. The outcomes of prediction of cancer regarding susceptibility, recurrence and survivability usually relates to (i) life expectancy, (ii) survivability, (iii) progression and (iv) treatment sensitivity [6].

1. *Prediction Of Cancer Susceptibility*

A case study was given in [6] which is relevant to estimation of breast cancer estimation through ANN. In case study, they developed a predictive model that distinguish between malignant mammographic findings and benign which is estimated by ten cross fold validation technique. It is developed from large number of hidden layers and small number of hidden nodes. The dataset used in this study has 48.774 mammographic findings with risk factors and tumour characteristics. These are fed as input to ANN model and trained this predictive model with mammography findings with tumour outcomes.

1. *Prediction Of Cancer Survivability*

This prediction gives health workers the idea of chances of survivability of patients after complete diagnosis with the help of ML techniques. Many predictive models are used for the evaluation of survivability. There is a requirement of huge data set for evaluation and prediction for robust results. In [6], SEER cancer database which has 162500 records with 16 key features and class variable named survivability is used to compare three productive models. These models are ANN, SVM and SSL. Five cross fold validation method is used to evaluate the performance of these models on the basis of accuracy. Using this technique, the accuracy rate for SVM is 51%, for SSL is 71% and for ANN is 65%. On comparing, SSL is better predictive model for survivability. There is a growing trend of using SSL predictive model for evaluation. There is another study regarding prediction of cancer survivability is also published. Data set contains gene’s raw expression data and 4 clinical variables namely age, sex, T\_stage and N\_stage is considered. The features of dataset are fed as input to ANN model. This input is used to train model and to produce robust results.

1. *Prediction Of Cancer Recurrence*

After diagnosis, the patients should be categorized into high risks and lower risks. It will give the idea of recurrence of cancer to health workers. A study regarding it is published in [6], in which three predictive models SVM, ANN and Cox-proportional hazard regression is used for prediction of breast cancer after diagnosis/treatment. The dataset used in this study has 14 features which is selected from total 193 features. These features are fed as input to our predictive models for our analysis. These models are evaluated on the basis of accuracy. The accuracy rate for SVM is 84.6%, for ANN is 81.4% and for Cox-proportional hazard regression is 72.6%. On the basis of accuracies, SVM has better performance regarding prediction of recurrence. Another study regarding prediction of cancer recurrence was also published which is based on oral squamous cell carcinoma. 86 patients were considered with their clinical, imaging and genomic features. These features are fed as input to predictive models. Out of 86 patients, 13 patients were identified with recurrence of diseases and other remaining patients were healthy. The predictive models are used in this are ANN, SVM, DT, RF and BN. Their performance is evaluated on the basis of accuracy, sensitivity, ROC curve and specificity using ten cross fold validation technique. On comparing on mentioned attributes, BN has better performance without any feature selection with clinical and imaging feature as input. Also, BN has better perform than other predictive models using CFS algorithm with genomic features as input.

IV. DISCUSSION

TABLE 1. PROS AND CONS OF VARIOUS ML ALGORITHMS.

|  |  |  |
| --- | --- | --- |
| **Algorithms** | **Advantages** | **Disadvantages** |
| Decision Tree | Can learn nonlinear relationships  efficient training algorithm | Individual trees are prone to over fitting |
| Support Vector Machines | Robust against over fitting  model non-linear decision boundaries | Complex structure of algorithm |
| Bayesian Network Learning | Able to directly calculate the probability that each drug is the best treatment | Requires extra statistical expertise |
| Artificial Neural Network | Tolerate noisy inputs  Parallel processing ability | Too many attributes result in over fitting |

The 10 times cross validation process is used to calculate high accuracy rate. It is basically used to classify algorithms on the basis of accuracy. The predictive accuracy of the model is computed from the testing set which provides an estimation. Accuracy is a measure of classifiers ability to produce the level of precise diagnosis. It is basically used to classify algorithms on the basis of accuracy. In [6], Wisconsin Diagnosis Breast Cancer dataset is used with the achievement of 97.38%. SVM has 97.14% accuracy and ANN algorithm has 96.71%. Table 1 is summarizing various machine learning algorithms on the basis of their advantages and disadvantages. The various advantage of machine learning includes continuous improvement as it is capable of learning from the huge volume of new data everyday , It also has the ability to automate various decision making task while its disadvantages includes data acquistion in which collecting data from survey results in incorrect data which leads to poor accuracy of models , It is time consuming as the time it takes to learn and process data increases with the increase in the volume of data.

CONCLUSION

Technology has brought severe change to practically every area and it is still evolving in all the fields of life with the continuous human efforts. A trial of machine learning in healthcare reveals how technology innovation can lead for more effective improvement of patent by helping the health care practitioner, it helps the volunteer in the healthcare field by reducing their work instead of replacing them. But volunteer need to provide accurate data to the machine, a false and inaccurate data to the machine could also make the circumstances worsen with the harsh consequences which can even lead to the death. Automatic cancer diagnosis is a significant real-world medical issue. Early detection of cancer is crucial for effective therapy. This research demonstrates how machine learning can be used to actual cancer diagnoses for a localized and systematic treatment.

In the future, data from many places throughout the world will be collected to construct a more accurate and general predictive model for cancer detection.

REFERENCES

**[1].** Katharine E. Henry, David N. Hager, Peter J.Pronovost, Suchi Saria, "A targeted real-time early warning score (TREWScore) for septic shock",Science Translational Medicine, vol 7, no 299, August 2015.

[2] Muhammad Javed Iqbal, Zeesham Javed, “Clinical application of artificial intelligence and machine learning in cancer diagnosis”, biomedical (BMC), may,2021.

[3] Ahan Chatterjee, “A Machine Learning Approach to prevent cancer “, The Neotia University, September 2020.

[4] Pushpanjali Gupta, Sam Fu Chaing, Prasan Kumar Sahoo, Suvendu kumar Mohapatra, jeng-fu you, Djeane Debora Onthoni,” Prediction of colon cancer stages and survival period with machine learning approach”,cancers,12 Dec,2019 .

[5] Tanzila Saba, Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges, Journal of Infection and Public Health, vol 13, no 9, pp 1274-1289,2020,

[6] Konstanine Kourou Dimitros Fotiadis,” Machine Learning Application in cancer prognosis and prediction”, Science Direct, November 2014 WHO report- Breast cancer prevention and control, January 2015.

[8] Pier Paolo Ippolito,” Understanding cancer using Machine Learning”, published in towards Data Science, January 2019