International Journal of Research in Engineering, Science and Management
Volume-1, Issue-10, October-2018
www.ijresm.com | ISSN (Online): 2581-5792

Cancer Detection and Classification Using AI

J. Dheeraj[1], S. Gurubharan[2]
[1,2]Student, Department of Computer Science, SRM Institute of Science and Technology, Chennai, India

Abstract—**A lot has been spoken about how precision medicine and, more particularly, how genetic testing is going to disrupt the way diseases like cancer are treated. But this is only partially happening due to the huge amount of manual work still required. Cancer is the general name for a group of more than 100 diseases. Although cancer includes different types of diseases, they all start because unwanted cells grow out of control. Without proper treatment, cancer can cause serious health problems and even loss of life. Early detection of cancer is necessary for its treatment. This paper presents a review of the detection methods for lung, breast, and brain cancers using Artificial Intelligence. The artificial intelligence techniques, such as support vector machine neural network, artificial neural network, fuzzy logic, and adaptive neuro-fuzzy inference system, with medical imaging like X-ray, ultrasound, magnetic resonance imaging, and computed tomography scan images are used. Imaging techniques are the most important approach for accurate diagnosis of human cancer. We investigated all these techniques to identify a method that can provide superior accuracy and determine the best medical images for use in each type of cancer.**

***Index Terms*— Medical Imaging, Artificial Intelligence Techniques, Lung Cancer, Breast Cancer, Brain Cancer**

I. Introduction

Cancer is the common name given to a group of related diseases. In all types of cancer, several body tissues start to divide into many parts, without stopping and spread around cells. Cancer can start at almost any place in our body, which is composed of approximately trillions of cells. Human tissues usually grow and divide to form new tissues as the human body needs them to. When cells age or become damaged, they die and are replaced by new cells. However, when cancer develops, this orderly tradition breaks down. As cells become aged, abnormal and damaged cells don't die and new cells form when they are not needed. These extra cells can divide without interruption and may form growths called tumors. Most types of cancers form solid tumors, which are composed of cell masses. Cancers of the blood, such as leukemia's, often do not form solid tumors as they form other types of tumors. Cancer tumors are malevolent in which they can spread into, or invade, adjacent cells. A number of cancer tissues from these tumors often break off and go to distant areas in the body. New tumors can spread to other areas away from the primary cancer growth through the blood or the lymph system. Benign tumors can often be removed, and in numerous instances, they do not go back and spread to other body parts. Cancer is a leading cause of disease worldwide. According to estimates from the International Agency for Research on Cancers, 14.1 million new cancer cases occurred and 8.2 million people died from cancer worldwide in 2012. AI Techniques: AI techniques are approaches that are utilized to produce and develop computer software programs. AI is an application that can re-create human perception. This application normally requires obtaining input to endow AI with analysis or dilemma solving, as well as the ability to categorize and identify objects. This paper describes various AI techniques, such as support vector machine (SVM) neural network, fuzzy models, artificial neural network (ANN), and K-nearest neighbor (K-NN). Research Methodology: Various intelligent techniques are utilized by researchers to help classify and segment medical image data to identify abnormalities within different areas of the body. This type of study is confined to the use of most of these techniques for classification and segmentation of medical image data. Medical imaging has developed into a crucial part of earlier diagnosis, detection, and treatment method of cancer through the years. Medical imaging is usually the first step to avoiding the spread of cancer via earlier detection and, in numerous cases, assists in the treatment or total elimination of cancer. CT imaging, MRI, mammography, ultrasound (US) imaging, X-ray imaging, and so on, are typical imaging modalities used for fighting cancer.

II. Classification of Cancers

Different types of AI algorithms are used to detect and classify different types of cancers. These techniques showed fluctuating accuracy across different years. This varying trend could be due to numerous factors, including network structure. In designing architecture for specific applications, the following selected parameters vary: network type, numbers of layers, number of nodes in hidden layers, activation function between layers, and the size of the data set used. Network generalization indicates how these networks are able to work with different data to decrease performance error to the lowest value.

A. *Breast Cancer*

Breast cancer is a malevolent tumor that starts in the tissues of the breast. This cancer can expand directly into neighboring areas or maybe distribute to distant parts of the body. The disease occurs mostly in females, but men can also develop this type of cancer. With recent functions involving examinations, considerable interest with regard to the utilization of adaptive strategies to assist detection and diagnosis of breast cancer among most cancers is concentrated on mammography. The particular mammography technique is a simple yet effective tool for prediction that involves breast cancer at an earlier stage. The Wisconsin Breast Cancer data source was offered by Dr. William H. Walberg, and the numerous AI techniques that are used by researchers and applied on WBCD database for prediction, detection, and classification of breast cancer are discussed in the succeeding paragraphs.

B. Brain Cancer

The central nervous system consists of the brain and the spinal cord. Brain tumor is probably the major driving force behind death occurring from cancer. Brain tumors are also known as gliomas. Two main types of brain tumor include brain cancer and tumors that start in the brain. Cancer cells can spread and enter healthy cells in the brain and spinal cord but not often affect other body parts. Secondary brain tumor is a more familiar type. The cancer begins within a different part of the body, like lung cancer or breast cancer, and spreads to the brain. This tumor is also known as a metastatic brain tumor. Brain cancer is likely curable and can be treated if detected at the initial stages. Without treatment, brain tumors can aggregate and cause death. Various techniques are used for obtaining images of human brain. These types of methods include X-ray, CT, electroencephalogram (EEG) signal, and MRI. These methods are used for diagnosis.

C. Lung Cancer

Lung cancer includes out of control development associated with unnatural tissues, which starts in a single or even in both the lungs. The unusual cells usually do not grow into healthy lung cells, instead separating rapidly and forming cancers. Major types of lung cancer generally involve non-small cell and small cell lung cancer. These cancers depend on how tissues appear under a microscope. Non-small cell lung cancer is more widespread compared with small-cell lung cancer. Lung cancer is probably the type of cancer that commonly leads to extremely high death rate. The most effective method of protection against lung cancer is early prognosis and diagnosis. Detection of lung cancer at an early stage is a complicated issue because of the construction of cancer tissues, in which almost all cells are overlapped. Besides being a crucial element in image processing, efficient identification of lung cancer at an initial stage is very much important.

III. RELATED WORKS

AI is not new, but there have been rapid advances in the field of machine learning and artificial intelligence in recent years. This has in part been plausible by developments in computing power and the huge volumes of digital data that are now generated. A wide range of applications of AI are now being explored with considerable public and private investment. The UK Government announced its goal to make the UK a world leader in AI and data technologies in its 2017 Industrial Strategy. In April 2018, a £1bn AI sector deal between UK Government and industry was announced, including £250 million towards AI research. AI is lauded as having the potential to help simplify important health challenges, such as meeting the care needs of an ageing population. Major information technology companies - including Google, Microsoft, and IBM - are investing in the development of AI for healthcare and research. The number of AI start-up companies has also been steadily increasing. There are many UK based start-up companies, some of which have been set up in collaboration with the UK universities and hospitals. Collaboration have been formed between NHS providers and AI developers such as IBM, Deep Mind, Babylon Health, and Ultromics. Such partnerships have attracted controversy, and wider concerns about AI have been the focus of several inquiries and initiatives within industry, and medical communities. AI has the potential to be used in planning and resource allotment in health and social care services. For example, the IBM Watson Care Manager system is being controlled and executed by Harrow Council with the aim of improving cost efficiency. It matches individual's interest with a care provider that meets their needs, within their allocated budget. It also designs individual care plans, and claims to offer insights for more effective use of health care management resources.

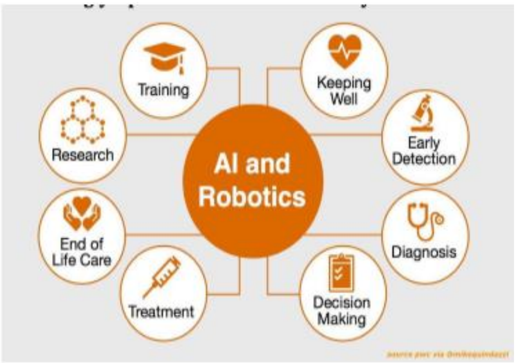

Fig. 1. AI in health care

IV. PROPOSED SYSTEM

Machine Learning, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference. Every learning process consists of two parts: (i) estimation of unknown dependencies in a system from a given data sample (ii) use of estimated dependencies to prognosis new outputs of the system. ML has also been proven an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an n-dimensional space for a given set of biological samples, using different methods and algorithms. There are two common types of ML methods known as (i) supervised learning and (ii) unsupervised learning. In supervised learning a labeled set of training data is used to estimate or relate the input data to the desired output. In contrast, under the unsupervised learning methods no labeled examples are provided and there is no picture of the output during the learning process. As a result, it is up to the learning scheme or model to find patterns or discover the groups of the input data. In supervised learning this process can be thought as a classification problem. The function of classification refers to a learning process that categorizes the data into a set of finite classes. Two other common ML functions are regression and clustering. In the case of regression problems, a learning function relates the data into a real-value variable. Subsequently, for each new sample the value of a predictive variable can be estimated, based on this procedure. Clustering is a common unsupervised task in which the user tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar traits that they share.

Suppose for example that we have collected medical records relevant to breast cancer and we try to predict if a tumor is malevolent or benign based on its size. The ML question would be referred to the estimation of the probability that the tumor is malignant or not (1 = Yes, 0 = No). Fig 2 shows the classification process of a tumor being malignant or not. The circled records depict any miss-classification of the type of a tumor produced by the process. Another common type of ML methods that have been widely applied is semi-supervised learning, which is a combination of supervised and unsupervised learning. It combines labeled and unlabeled data in order to develop an accurate learning model. Usually, this method of learning is used when there are more number of unlabeled data-sets than labeled. When applying a Machine Learning method, data samples constitute the basic components. Every sample is described with several features and every feature consists of various types of values. Furthermore, knowing in advance the specific type of data being used allows the right selection of tools and techniques that can be utilized for their analysis. Some data-related issues refer to the quality of the data and the preprocessing steps to make them more suitable for Machine Learning. Data quality issues include the presence of noise, outliers, missing or duplicate data and data that is biased and unrepresentative. When improving the data quality, the quality of the resulting analysis is also improved. In addition, in order to make the raw data more suitable for further analysis, preprocessing steps should be applied, on the raw data, that focus on the modification of the data. A number of different techniques and strategies exist, relevant to data preprocessing that focus on modifying the data for better fitting in a specific Machine Learning method. Among these techniques some of the most important methods include (i) dimensionality reduction (ii) feature selection and (iii) feature extraction. There are many advantages regarding the dimensionality reduction when the datasets have a large number of features. ML algorithms work better when the dimensionality is lesser. Additionally, the reduction of dimensionality can eliminate irrelevant features, reduce noise and can produce more robust learning models due to the involvement of lesser number of features. In general, the dimensionality reduction by choosing new features which are a subset of the old ones is known as feature selection. Three main methods exist for feature selection namely embedded, filter and wrapper approaches. In the case of feature extraction, a new set of features can be created from the initial set that captures all the significant information in a dataset. The creation of new sets of features allows for gathering the described advantages of dimensionality reduction. However, the application of feature selection techniques may result in specific fluctuations concerning the creation of prognosis feature lists. Several studies discuss the phenomenon of lack of agreement between the predictive gene lists discovered by different groups, the need of thousands of samples in order to achieve the desired outcomes, the lack of biological interpretation of predictive signatures and the dangers of information leak recorded in studies.

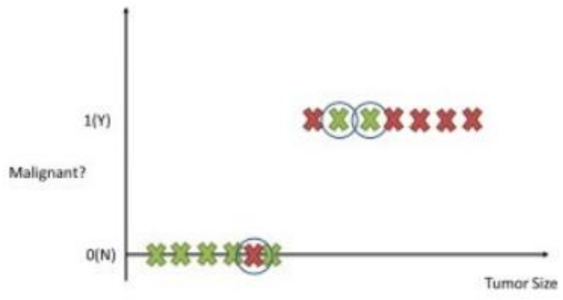

Fig. 2. Malignant

The main objective of Machine Learning techniques is to produce a model which can be used to perform classification, prediction, estimation or any other similar task. The most frequent task in learning process is classification. As mentioned previously, this learning function classifies the data set into one of several predefined classes. When a classification model is developed, by means of Machine Learning techniques, training and generalization errors can be produced. The former refers to classification errors on the training data while the latter on the expected errors on testing data. A good classification model should fit the training set appropriately and accurately classify all the instances. If the test error rates of a model begins to increase even though the training error rates decrease then the phenomenon of model over fitting occurs. This situation is related to model complexity meaning that the training errors of a model can be reduced if the model complexity increases. Obviously, the ideal complexity of a model not susceptible to over fitting is the one that produces the lowest generalization error. A formal method for analyzing the expected generalization error of a learning algorithm is the bias–variance decomposition. The bias component of a particular learning algorithm measures the error rate of that algorithm. Additionally, a second source of error over all possible training sets of given size and all possible test sets is called variance of the learning method. The overall expected error of a classification model is constituted of the sum of bias and variance, namely the bias–variance decomposition. Once a classification model is obtained using one or more Machine Learning techniques, it is important to estimate the classifier's performance. The performance analysis of each proposed model is measured in terms of sensitivity, specificity, accuracy and area under the curve. Sensitivity is defined as the proportion of true positives that are correctly observed by the user, whereas specificity is given by the proportion of true negatives that are correctly identified. The quantitative metrics of accuracy and area under the curve are used for assessing the overall performance of a classifier. Specifically, accuracy is a measure related to the total number of correct prognosis. On the contrary, area under the curve is a measure of the model's performance which is based on the ROC curve that plots the tradeoffs between sensitivity and 1-specificity. The predictive accuracy of the model is calculated from the testing set which provides an estimation of the generalization errors. In order to obtain reliable results regarding the predicting performance of a model, training and testing samples should be sufficiently large and independent while the labels of the testing sets should be familiar. Among the most commonly used approaches for evaluating the performance of a classifier by splitting the initial labeled data into subsets are: (i) Holdout Procedure, (ii) Random Sampling, (iii) Cross-Validation and (iv) Boot strap. In the Holdout procedure, the data samples are partitioned into two separate sets, namely the training and the test sets. A classification model is then computed from the training set while its performance is estimated on the test set. Random sampling is a similar method to the Holdout method. In this case, in order to obtain better estimate of the accuracy, the Holdout method is repeated several times, choosing the training and test instances randomly. In the third procedure, namely cross-validation, each sample is used the same number of times for training and only once for testing. As a result, the original data set is gone through successfully both in the training and in the test set. The accuracy results are calculated as the average of all the various validation cycles. In the last approach, bootstrap, the samples are separated with replacement into training and test sets, that is, they are placed again into the entire data set after they have been selected for training.

V. Conclusion

In this review, we discussed the concepts of Machine Learning while we outlined their application in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised Machine Learning methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques and procedure for feature selection and classification can provide promising tools for inference in the cancer domain.

Acknowledgement

This research was supported by our teachers, friends, parents supervised by the SRM Institute of Science and Technology.

References

[1] Ada, R. K. (2013). Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier: IJAIEM.

[2] Al-Daoud, E. (2010). Cancer diagnosis using modified fuzzy network. Universal J. Comput. Sci. & Engg. Technol, 1(2), 73-78.

[3] Al-Naami, B., Mallouh, M. A., & Hafez, E. A. (2014). Performance Comparison of Adaptive Neural Networks and Adaptive NeuroFuzzy Inference System in Brain Cancer Classification. JJMIE, 8(5).

[4] Al-Timemy, A. H., Al-Naima, F. M., & Qaeeb, N. H. (2009). Probabilistic neural network for breast biopsy classification. Paper presented at the Developments in e Systems Engineering (DESE), 2009 Second International Conference.

[5] Anandgaonkar, G. P., & Sable, G. S. (2013). Detection and Identification of Brain Tumor in Brain MR Images Using Fuzzy CMeans Segmentation.