Early Diagnosis and Prediction of Recurrent Cancer Occurrence in a Patient Using Machine Learning

Swarn Avinash Kumar[1], Harsh Kumar[2], Srinivasa Rao Swarna[3], Vishal Dutt[4]

[1]IIIT Allahabad
[2]People's Friendship University of Russia (RUDN)
[3]Tata Consultancy Services, Edison, NJ
[4]Department of Computer Science, Aryabhatta College, Ajmer, India

e-mail: kswarnavinash@gmail.com[1], harshkumaronn8@gmail.com[2], swarna.prince@gmail.com[3], vishaldutt53@gmail.com[4]

Abstract: Machine learning is the way to implement many kinds of research applications which will create a challenge to implement the novel tasks. Cancer is the most important research component in the medical field which requires machine learning to look over the solution. There are a lot more chances of the occurrence of cancer to a specific person even after the recurrent sessions of the treatment and there is no way to recognize with the clinal knowledge. We need a hand of an expert system to analyze the patients' present condition and need to recognize the better path for the patient for his or her treatment or the life span. Machine learning implementation in the medical domain mostly work on the classification mechanisms in the initial stage of the implementation and we need to work out on implement the ensemble models like the random forest, AdaBoost mechanisms which will give the challenging training methods for the model and the models which are being trained using the ensemble methods will give the accurate results related to any kind of the disease treatments if the concept of implementation of the machine learning is in the initial stage of the implementation. The machine learning models like Random Forest, AdaBoost, SVM, Decision trees gave some respectable results concerning the identification of the patient's condition who got treatment for cancer and in the post-treatment stage. Among all the models, the random forest gave the highest accuracy of predicting the cancer post-treatment with 93% and SVM got some least among the list with 82% of accuracy in early identification.

1. INTRODUCTION

Machine learning and the medical domain are the best pair for the implementation of the research components. The different research components available in the medical domain will give the best feel to learn and implement the better models all the time and there is a gap in the implementation of the machine learning models which have some training and test case problems which make the models week in implementation. The machine learning models like support vector machine, decision trees, and XGBoost will create a small impact in predicting the problem caused for the patient in the future. But we need to look at the occurrence of cancer in the patient in the post-treatment stage. PTS (Post Treatment Stage) is the significant data that can have the impact of the treatment and medication on the patient which can be further used for the prediction implementation. For every instance of the data relation, there will be an internal relation between the hidden components. For an instance consider a patient who got treatment for cancer (Any Kind), and after the treatment, what are the food habits of the patients and the timeline of the medication are some hidden things. When the patient is using the medication and what is the timeline the patient is taking the medication is the other internal instance which can be considered as the hidden component in the research implementation. The hidden components may affect the implementation of the PTS using machine learning models. The PTS is the most important factor we need to work on. The PTS stage is having different hidden stages. These stages are meant to be most important for the identification of the feature extraction and feature selection.

The dimensionality reduction is the concept of the feature extraction and feature selection in which we use the concept of PCA (Principal Component Analysis). The PCA is the most effective factor which handles the probability of considering the specific feature in the modeling. Because of the more hidden rules and the features we need to classify the inner relation in the initial stage of the modeling and in the next part of the modeling we need to classify the better analysis. The better analysis deals with the appropriate modeling when the features are of a good standard. The extraction is happening concerning the relativity between the two features. For an instance we can consider the blood pressure (BP), CBP (Complete blood picture) are the important combinations. But in some cases, we can’t find the accurate relation between BP and CBP. They both quite contradict each other.

Machine learning models like random forest will work on the following modeling diagram using which we are trying to classify the implementation of the analysis on the cancer patients. There is no isolation of the individual cancer types.

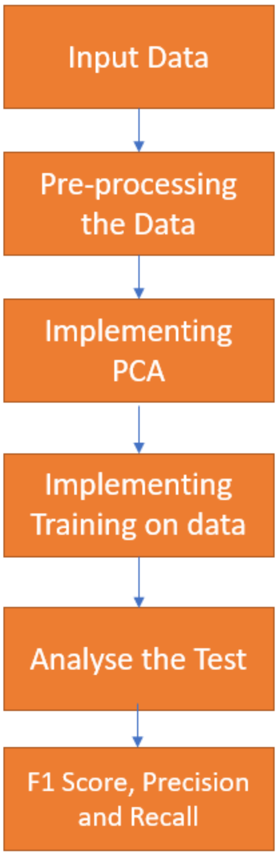


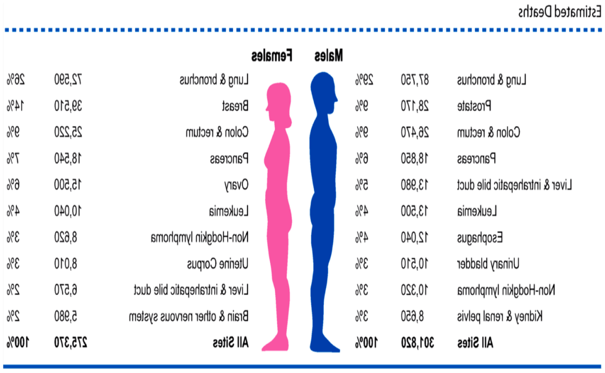
Figure 1: Processing of the data using PCA

The PCA will help the implementation in a better way with feature extraction and selection. The concept of feature extraction can be performed using backward elimination and forward chaining. The concept of backward elimination can help to identify the best features for the modeling. All the features in the dataset are not useful. But there will be few features that will be the most important for modeling. The training accuracy will be more when there is any implementation like backward elimination of forwarding chaining. These mechanisms will take P and SL values are the main base and if the features which are having a P-value greater than 0.5 then that value will be eliminated. The prediction estimates the criticalness of the patient's condition after the cancer treatment. For an instance consider the patient with cancer and that patient undergone treatment like chemotherapy. After some stages of treatment, we can see that the infected area got cured. But the reality is there is no complete cure for the infected region, instead, there is a high chance of getting it back with the other symptoms. The proposed system in the current work implements an efficient algorithm that can predict the patient’s condition post-treatment.

The lateral part of the sections deals with the cases of the cancer stages, next with the methodology of machine learning implementation, next explains the literature review on the current scenarios, next deals with the proposed methodology, and conclude with the results.

2. DIFFERENT TYPES OF PROBLEMS IN CANCER MACHINE LEARNING NEED FOCUS

There are different types of cancer problems we need to focus on. As mentioned earlier there are different hidden insights of the data which is having the interrelation between the variables like features. For an instance consider that there is an instance of identifying the patient's condition before and after the treatment. Figure 2 describes different types of cancers and deaths in both the gender



Estimated Deaths

Males			Females		
Lung & bronchus	87,750	29%	Lung & bronchus	72,590	26%
Prostate	28,170	9%	Breast	39,510	14%
Colon & rectum	26,470	9%	Colon & rectum	25,220	9%
Pancreas	18,850	6%	Pancreas	18,540	7%
Liver & intrahepatic bile duct	13,980	5%	Ovary	15,500	6%
Leukemia	13,500	4%	Leukemia	10,040	4%
Esophagus	12,040	4%	Non-Hodgkin lymphoma	8,620	3%
Urinary bladder	10,510	3%	Uterine Corpus	8,010	3%
Non-Hodgkin lymphoma	10,320	3%	Liver & intrahepatic bile duct	6,570	2%
Kidney & renal pelvis	8,650	3%	Brain & other nervous system	5,980	2%
All Sites	**301,820**	**100%**	**All Sites**	**275,370**	**100%**

Figure 2: Types of cancers and deaths in both the genders

Considering figure 2 we can identify that in males and females both lung and bronchus cancer symptoms and deaths are high compared to other types. In the lease, we have kidney failure. But there is an interrelation between the variables where we need to figure out the symptoms of every case. There will some common symptoms where machine learning needs to focus on. Figure 3 consists of the common symptoms in all the cancer types and those will feed the machine with the required information to map the relativity in the data.

The relativity needs to be considered while doing any kind of machine learning medical implementation. The implementation lies with the data and the factors we are considering for a better implementation.

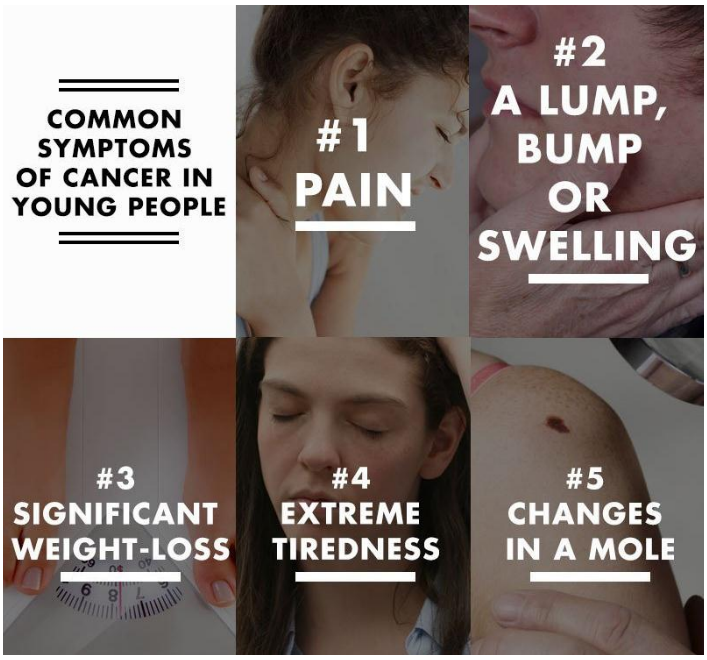

COMMON SYMPTOMS OF CANCER IN YOUNG PEOPLE
#1 PAIN
#2 A LUMP, BUMP OR SWELLING
#3 SIGNIFICANT WEIGHT-LOSS
#4 EXTREME TIREDNESS
#5 CHANGES IN A MOLE

Figure 3: Common symptoms of cancer.

Based on figure 3 we can estimate that there is a high chance of misunderstanding that can happen while deciding the type of cancer the patient is having with these 5 common symptoms. But the machine learning model needs to focus on gathering accurate data related to the patients in these top 5 symptoms. Table 1 describes the machine learning requirements of the better designing process.

Table 1: Machine Learning Requirements in the Cancer Prediction

Feature	**Description**
Blood Pressure	Consider the BP report for every 15 days
Diabetic	Diabetic values will be changing from time to time. Its dynamic data
Weight	Patients weight will lose in a shorter period than expected (note the time stamp)
Dizziness	Extreme dizziness within time stamps (Note the duration)
Skin Rashes	The patient needs to observe their skin for the extreme amount of moles and rashes
Pain	Excessive body pains (Note the duration of pain)
Swelling	Organs swelling. (Need to note the volume)

In table 1 we need can find the variables which are most useful for a better understanding. We need to know the hidden things which are a lot of mean for the prediction model design and implementation. In the survey conducted related to the hidden features mentioned above, the following table2 contains the features which are having the highest rating for maintaining in the prediction model design and implementation.

The rating is designed based on the manual survey. The survey questions consist of the answers in the form of rating from 1 to 5. 1 is the least and 5 is the highest. Table 2 contains the hidden features which are having the highest rating.

Feature	Rating
Blood Pressure	4
Diabetic	5
Dizziness	3
Pain	3
Weight	4

The features which got a rating greater than or equal to 3 are considered are the best-hidden features. This survey is conducted manually without any predictions.

3. MACHINE LEARNING METHODOLOGY

The machine learning implementation is the most important thing the disease treatment identification and there is a chance of implementation of machine learning with a small part then need to go for the larger step. As mentioned earlier there are two things we need to do while considering cancer as the main base for medical processing.

i. Pre-Treatment analysis

In the pre-treatment analysis (Pr-TA) the major factor is where we will get all the information related to table 1. All the information related to the patient we need to gather and make it as a dataset. The corresponding cases related to the same disease need to be identified with the same features. The features which are considered in the Pr-TA session need to be carried forward to the Po-TA. This can be considered later. This session will take the necessary measures to identify the severity of the patient concerning the disease. Including the features mentioned in table 1, we need to consider all of them and need to monitor every same thing after the treatment.

ii. Post – Treatment analysis

In the post-treatment analysis (Po – TA) the major factor is to maintain the same features to the prediction model design. The prediction model using the random forest or any other modelling will be based on this kind of final result only. The concept which this article is speaking about is the recurrent occurrence of the disease after the treatment. For that analysis, we need to consider the post disease treatment mechanisms which are handling the major part of the hidden features.

iii. Random Forest

As it is an ensemble model we need to figure out the rules we need to assign for an individual decision tree. As the random forest is the most important and accurate thing that happened in this research, we need to focus on the pairing of the features which make the best prediction model set. The prediction model which gives the highest accuracy will have the best ground truth mechanism like confusion matrix. The confusion matrix will state the positivity and negativity of the result obtained. This case has the highest positivity in the result with random forest.

iv. Decision Trees

As we are implementing the concept with random forest, but for the smaller pairs and the for the small surveyed features we can use the decision trees as they work easily for the single type of rule which is not more complex than the random forest implementation.

v. Support Vector Machines

We need to check the least possible case of the prediction model design as there are lot more chances of the prediction model failure with not using the proper prediction model. The modelling and the sampling for the training and testing is the most important factor which works on defining the type of result we occur using the confusion matrix. Support vector machine works with considering the one feature as the independent variable and classifies whether the feature is needed for the level 2 modelling or not.

4. LITERATURE REVIEW

The literature review is a short note for the existing implementation we have. We have different diseases to discuss. For an instance consider the diseases like dementia, heart disease and other congenital disorders which is having the highest chance of failure rate if there is no proper identification of the symptoms after treatment. For dementia, we can't do anything because of not having the proper support from any medical organization. But there is a chance of implementation of machine learning in heart disease prediction. In this, we can gather the data related to heart disease patients who got treatment. But using the machine learning models we need to track the performance of the patient with in the time intervals. The time intervals are the most important thing we need to focus on. There are many chances of failure of the problem-solving approach in machine learning. If there is no proper data to handle and train, we can lose the model.

Some of the cases like deep learning implementation, which consists of the hidden layers, where we need to implement using the activation functions. The activation functions will make the work done for the researcher without performing feature selection and extraction. The reason for using the traditional method for the modelling is to analyse the best factors from the data. Inter-relativity is the most important feature. For every dataset, there will be interrelated data, where we need to get the insights of the data. There will be a hidden relationship among the features and no article will speak about the hidden relations.

Different researchers are working on different medical research problems where we need to work on both image and textual data. The data which have in the recent research is on different neurological disorders.

5. PROPOSED METHODOLOGY

The proposed methodology consists of a two-layer implementation. The first later will be implemented using the Decision trees and SVM. The second layer will be implemented using the random forest, XGBoost and AdaBoost mechanisms.

i. Layer 1 (Pr-TA)

The Pre-Treatment analysis is the major factor which will be implemented using the decision trees. The pre factors will have some valuable information. For an instance consider the patient with some information which is the small and tiny thing. This may be useful for better prediction later.

ii. Layer 2 (Po – TA)

The data we gathered from the pre-treatment analysis will be taken forward to the post-treatment analysis which is having implementation with DNN, Random Forest. DNN is having the highest accuracy among the other modelling.

The algorithms of the Pr – TA and Po- TA are as follows

```
Pr- TA (Features, Timestamp, Result){
Check the features -> Timestamp
Pre-processing the data
{
PCA Implementation{
Extract the features
}
Implement the confusion matrix
}
Return result
}
Po- TA (Pr- TA, Features, Timestamp, Result){
Check the features -> Timestamp
Pre-processing the data
{
PCA Implementation{
Extract the features
}
Implementation of DNN, RF, AdaBoost
Implement the confusion matrix
}
Return result
}
```

6. RESULT AND CONCLUSION

As mentioned in the proposed system the DNN will take the charge of defining the variables which are effecting the modelling. The modelling used in this approach is the most useful. It can be a genetic model which can be used for any kind of cancer problems. Figure 4 defines the result of the prediction model design. The design of the algorithm will help the researchers to implement an efficient way of two-layer implementation of the disease prediction. The disease prediction models are the most important factors where we have the different models to present. The DNN is the deep neural network implementation where we can get the access to the dataset and implement the hidden layers. The hidden layers are the most important factors with activation functions which lead to the prediction analysis in a better manner.

Multi Layer Perceptron (MLP) is the other approach in which we gather the nodes in the neural network and perform the activation function in a better manner.

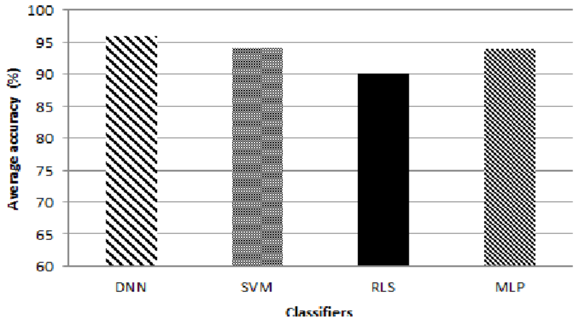


Figure 4: Result of the prediction models.

The DNN model gave the highest accuracy in all aspects in understanding the requirements of the modelling and there is a huge requirement on understanding the importance of the DNN, RF and DT for predicting the recurrent occurrence of cancer.

7. REFERENCES

[1] “Deep Radiomic Nalyssis Based on modelling Information Flow in Convolutional Neural Networks” Ahmad Chaddad et al, Volume 7, 2019, IEEE Access

[2] “Deep Radiomic Analysis of MRI related to Alzheimer’s disease” ahmad chaddad et al, Volume 6, 2018, IEEE Access

[3] MCADNNet: Recognizing Stages of Cognitive Impariment through efficient convolutional fMRI and MRI neural network topology models”, Saman Sarraf et al, Volume 4,2016, IEEE Access

[4] " Ensembles of patch based classifiers for diagnosis of Alzheimer's Diseases", Sa,suddin Ahmed et al, Volume 7, 2019, IEEE Access

[5] "Transfer learning with intelligent training data selection for prediction of Alzheimer's Disease" Naimul Mefraz Khan et al, Volume 7, 2019, IEEE Access

[6] "Hippocampus localization using two stage ensemble hough convolutional neural network", Abol Basher et al, Volume 7, 2019, IEEE Access

[7] "Deep Learning framework for alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM", Chiyu Feng et al, Volume 7, 2019, IEEE Access

[8] "Big Data Visualization in Cardiology—A Systematic Review and Future Directions" SHAH NAZIR et al, Volume 7, 2019

[9] "In Search of Big Medical Data Integration Solutions - A Comprehensive Survey" HOUSSEIN DHAYNE et al, Volume 7,2019

[10] "Radiogenomics for Precision Medicine With a Big Data Analytics Perspective" Andreas S. Panayides et al, Volume 23, No:5, September 2019

[11] "Intelligent Analysis of Medical Big Data Based on Deep Learning" by HANQING SUN et al, Volume 7, 2019, special section on deep learning algorithms for internet of medical things

[12] "Health Big Data Classification Using Improved Radial Basis Function Neural Network and Nearest Neighbor Propagation Algorithm" by CONGSHI JIANG et al, Volume 7,