Cancer Prediction using Multimodel Analysis

Akshat Katiyar
Department of Computer Science
PSG College of Technology
Coimbatore, India

Abstract— **Machine learning and artificial intelligence are the state-of-the-art technologies in the world at the moment. Amidst every industry keenly focusing on adapting these technologies onto their respective domains, learning machine learning opens up a plethora of opportunities to develop advanced machinery and applications related to face recognition, cybersecurity, medicine. As a team, we were intrigued by the use cases of this technology and its potential in creating an efficient environment altogether. Surfing on the Internet for its information and we ended up learning about the existence of deep learning. Getting acquainted with its initiatory had us intrigued even more; which is the reason why we took up this project. The role of multimodal analysis in cancer detection is a game changer because it defeats the purpose of manually inspecting the slides on the biopsies and checking for the gene-gene interaction, which takes up a humongous amount of time and is predominant in the world. This project not only eliminates the need to inspect the slides manually but also increases the accuracy of the measurement to a great extent. This drawback of manual inspection of the tissue slides by the lab technicians and the doctors invigorated us to innovate and execute an application that is critically needed by mainstream medical domain at the very moment. In case of Gene expression data, the need for finding the correlation between genes is reduced as KNN relates the most dominant gene for a particular user. Hence Multimodal analysis is used to gather the features of all the modals and creating a single modal which is able to predict effectively. In this case KNN is suited for data points while CNN is the game changer when it comes to images also SVM is highly reliable in case of small dataset as it classifies and predict at the same time.**

Keywords—CNN; SVM; Cancer; KNN; Ensemble Learning.

I. INTRODUCTION

The purpose of this project is to analyze and determine whether the given biopsy report of a person has been affected by cancer tumors or not by making use of Convolutional Neural Networks (CNN) and various other machine learning models. It is observed that the aforementioned method can deliver more efficient and optimized performance when compared to mid-range/manual performers on the same dataset. The system is aimed at performing well with dataset containing datapoints, images and also with gene expressions. In the existing system, these processes have been done using visual inspection of slides prepared from biopsies. Researchers have begun applying deep learning methods to cancer diagnostics. Hence, using machine learning techniques such as Artificial Neural Networks (ANN) has become necessary and a critical priority. Unlike traditional machine learning models, CNN learns from the fed dataset by constructing an input-output mapping for the problem at hand. Neural networks are more noise-tolerant and flexible when compared to traditional statistical models. In this project we are having 3 kinds of dataset related to breast cancer (images, gene expression and clinical data). The Images are treated with CNN and clinical and the gene data is treated with SVM and KNN respectively. The combined dataset of images (after compression), gene and the clinical data is given as input to the SVM.

Analysis of biopsy is one of the most critical and prolonging phases in the course of cancer detection. The existing system involves medical professionals inspecting the slides of biopsies and determining the presence of cells affected by mitosis by only manually analysing themselves. There is an absence of a technological aid to this process. As the manual inspection of slides is an infinitesimal work to be done, human errors tend to occur. For instance, the tumours present at the corners and edges of the slides could be left unchecked due to the splitting of slides and lack of observation. The discrepancies in the existing case are unlikely to be taken lightly of, as cancer detection is of a critical priority. The problem statement of this project is to analyse the given patient data(gene, images, clinical) and predict whether he/she has cancer or not using different kinds of models and integrating them in one unified model, which lays the foundation of multimodal analysis. The models used are KNN CNN and SVM. The data can be in any form either datapoints or images or gene expressions. The major problem we are solving here is unifying the given data from various sources into a single data file which will give a better prediction accuracy.

The goal of this project is to integrate various models that are available in analysing the cancer datapoints, images and gene expressions.

- To scan the given breast cancer biopsy dataset (categorized into 2 classes: Normal and Affected) through image processing. To analyse the scanned images through 3 layers of CNN: Input Layer, Hidden Layer, and Output Layer.
- To analyse the gene expression of a given patient by finding the dominant genes and getting the interaction between various genes.(SVM).The next objective is to make a cluster which causes cancer

and genes which doesn't cause the cancer. This is done using KNN.

- The major challenge to be faced in this is the data integration. No dataset consists of all the three types of data. Formulating a dataset with all the three types of data is done by filling missing values using some feature extraction method. The feature extraction should be done for all three types of data and the common feature should be used to make the main dataset.

This Research will be worthy of use to the medical professionals who uses manual techniques to examine the patient and detect the presence of cancer and classify them. Since, the project is implemented using multiple dataset from various sources ,it can provide a result for any single type of data available. It can be done either for datapoints or images or gene expressions. Also, the combinations of any of the above. The scope of the project not only reduces the complexity of the cancer analysis but also provides the examiner with high accuracy of their prediction.

II. RELATED WORK

A. *DESIGN*

The system is designed using convolutional neural network .In machine learning, a convolutional neural network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural networks that has successfully been applied to analysing visual imagery.

The layers of our system:

- The convolutional layer
- The pooling layer
- The output layer

The Convolution Layer extracts certain features from the image by defining the weight matrix. This weight runs across the image such that all the pixels are covered at least once, to give a convolved output. In our case, we use a 3*3 matrix.

The Pooling Layer reduce the number of trainable parameters. It is desired to periodically introduce pooling layers between subsequent convolution layers. Pooling is done for the sole purpose of reducing the spatial size of the image. Pooling is done independently on each depth dimension, therefore the depth of the image remains unchanged. The form of pooling layer applied here is the max pooling.

After layers of convolution and pooling, we output the result in the form of a class. The convolution and pooling layers would only be able to extract features and reduce the number of parameters from the original images. However, to generate the final output we need to apply a fully connected layer to generate an output equal to the number of classes we need the output layer has a loss function like categorical cross-entropy, to compute the error in prediction. The loss function used here is binary cross entropy since we output the binary output. The optimizer function used is Adam's function. The activation function for the input layer is ReLu and for the output layer is sigmoid function.

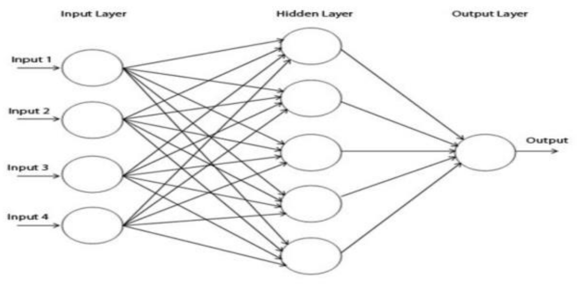


Fig 1. Network Structure

Reason for using SVM and KNN SVM is a model which is capable of doing both classification and regression. Especially the Non-linear dataset. On-linear SVM means that the boundary of that the algorithm need not be a straight line. Since , we are having a complex and convolutional dataset SVM can handle it easily. But one of the disadvantage of the SVM is that the training time of is much longer since, it has intense computations. KNN is one of the algorithm which does not make the training data points to do any generalization. It has very minimal training phase or it is not explicit training as such. This means the training phase is pretty fast . Lack of generalization means that KNN keeps all the training data. More exactly, all the training data is needed during the testing phase. As in our case the gene expression data seems to be completely random and also dependent to some extent hence KNN suits the purpose when gene-gene interaction comes into play.

B. *GENE EXPRESSION*

Gene expression is the process of working with the gene data and using them in advantage of their classification. The main aim is to find the attributes in the dataset which is affecting the cancer producing mitosis in the cells of the patients. Our aim is also to find the dependency that exist between the attributes. Since there are large number of attributes in the gene expression dataset the complexity of the project becomes high. So, we have only included specific attributes which have been creating the cancer in the past.

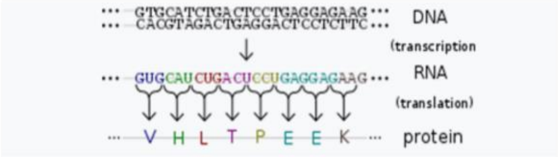


Fig 2. Gene Structure

III. DATA ANALYSIS

The purposed method for developing the system consists of mainly two main steps. Firstly, data is collected from web. Secondly, relevant features are selected. Then, an CNN with the above specified layers is designed and a suitable algorithm yielding best accuracy is chosen to detect the cancer. Then a classification is performed using the gene expression dataset.

A. Data Source

This project basically attempts to predict whether a particular patient has cancer or not. So, it is necessary to have a trusted source having relevant and necessary data required for the prediction. We will be using the data from https://becominghuman.ai website as the primary source of data. This website contains all the images of patients' breast histology report. The images are categorized as normal (without cancer) and affected (with cancer).

B. Selection of Dataset

The cancer disease can be traced in various organs of the body. There are many different parameters which can vary differently for each organ. In this case, we are applying our rules for breast cancer images. So, this project performs analysis and prediction on only the companies that fall in the IT sector.

C. CNN Design and Training

- Designing part of this CNN is difficult because all the iterating conditions are identified only based upon the Trail and Error method we can't find those values directly. And training the code is quiet easy because we have data set ready, we give the data set as input and on processing the data it will get trained.
- We basically need two groups of Data-set one is Training Data-set, the other is Testing Dataset. Training Data-set to train the model and Testing Data-set is to test the accuracy of prediction. Both these Data-set should contain both the cancer affected image and the normal images. We use 70% of the images for training the data and the remaining 30% of the images for testing the data.
- We normalize the data by applying Max pooling (i.e.), picking up the highest valued pixel from the specific region of interest. The other method which we apply here is flattening of the data. It converts the 2 dimensional image matrix into one dimensional vector.

The various modules involved as part of the project are:

- Making basic Modal for each technique used such as SVM, CNN,KNN etc.
- Aggregation all models in one big Multimodal.
- Mapping the gene expression of one patient with his clinical data.

IV. SYSTEM IMPLEMENTATION

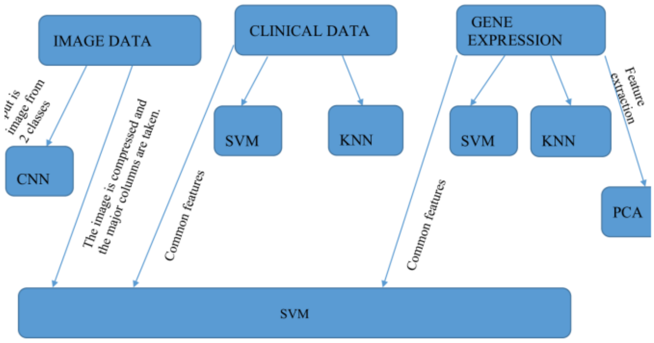


Fig 3. System Implementation

In the system implementation, the image data is classified with CNN separately and with SVM as a whole with clinical data and the gene expression. The reason for using the CNN is that it uses a feedforward mechanism,(the output of one layer is the input to the next layer) and is also used for classifying the images. Since, our input is images we have gone for the CNN. The clinical and the gene expression is classified using both, the SVM and KNN and as a whole with the SVM. The reason for using them is that, SVM is a model which is capable of doing both classification and regression. Especially the Non-linear dataset. On-linear SVM means that the boundary of that the algorithm need not be a straight line. Since , we are having a complex and convolutional dataset SVM can handle it easily. But one of the disadvantage of the SVM is that the training time of is much longer since, it has intenseness computations. KNN is one of the algorithm which does not make the training data points to do any generalization. It has very minimal training phase or it is not explicit training as such. This means the training phase is pretty fast . Lack of generalization means that KNN keeps all the training data. More exactly, all the training data is needed during the testing phase. As in our case the gene expression data seems to be completely random and also dependent to some extent hence KNN suits the purpose when gene-gene interaction comes into play. Since, our model is using different types of data from varied sources our aim is narrowed in choosing the classifier that will enable us improve the accuracy along with integrating the data types.

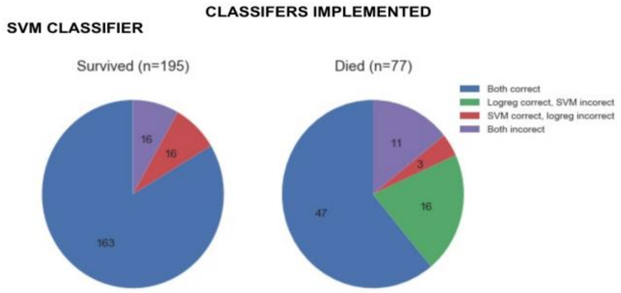
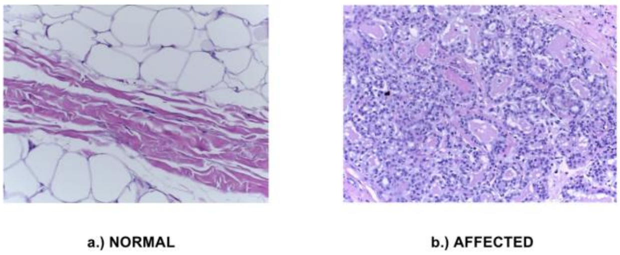


Fig 4. SVM CLASSIFIER PREDICTION

V. TESTING AND RESULT

Test set consists of 30% of the data-set. Once the model is built, the model is validated using test data. Binary cross entropy is used for finding the error.

ACCURACY: The accuracy of the model is 82.2%.


a.) NORMAL

b.) AFFECTED

Fig 5. Normal Image & Cancer Image

Gene types and their role in causing Cancer

- **al157500**: RUVBL1 was reported in metastatic prostate cancer cells
- **al157502**: Widely expressed in various human tissues - thymus and testis appear to express the highest levels.
- **contig19951**: RASSF1A and RASSF1C are expressed in heart, brain, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, small intestine, colon, peripheral blood leukocytes. RASSF1B is expressed predominantly in cells of hematopoietic origin.
- **contig20749**_rc:The RASSF family of tumour suppressor genes (TSG) encode Ras superfamily effector proteins that, amongst other functions, mediate some of the growth inhibitory functions of Ras protein
- **contig29982**_rc: This protein is an ubiquitously expressed nuclear protein and it belongs to a highly conserved subfamily of WD-repeat proteins. It is found among several proteins that bind directly to retinoblastoma protein, which regulates cell proliferation.
- **contig37376**_rc: RBX1 gene is evolutionarily conserved from plants to mammals with multiple family members in each species
- **contig44916**_rc: RCHY1 protein was found primarily in the cytoplasm and membrane and a small portion in the nucleus of malignant cells.
- **contig45645**_rc: gastrointestinal tract, stomach, small intestine, colon, and pancreas.
- **contig46501**_rc: The RecQ4 gene is predominantly expressed in thymus and testis and at low levels in other organs such as heart, brain, placenta, pancreas, small intestine, and colon, indicating that the expression of RecQ4 gene is somewhat tissue-specific.
- **contig47230**_rc: RBM5 is located at the human chromosomal locus
- **contig53047**_rc: Found in human skin raft culture. The origin, description and the expression of rest of the cancer tissues is not known.

TYPES OF DATA TAKEN	CLASSIFIER APPLIED	ACCURACY
Images	CNN	64 %
Clinical Data	SVM	88 %
Gene Data	SVM	73.2%
Gene Data	Logistic Reg	75%
Clinical Data	KNN	82.3%
Gene Data	KNN	72.7%
Clinical + Gene	Logistic Reg	83.8%
Clinical + Gene	SVM	84%
Clinical + Gene + Images	SVM	88%

Table 1. Accuracy of Ensemble model

VI. CONCLUSION

This model is applied for sample datapoints, images and gene expressions. Further detection can be effectively and efficiently done for other classes by using other different algorithms From our inferences that we have gained from this project, the SVM classifier has the highest accuracy. The reason for it having the higher accuracy is that it classifies and predicts at the same time. Analysis using the gene expression gave us more accuracy than dealing with the clinical data. The limitations that we encountered while producing the system , is making the image and the gene expression compatible. The IDs and the columns were not matching for all the images. So, it is done for sample images from the dataset and the accuracy of 88% is achieved.

REFERENCES

[1] Object-level inter-observer agreement and comparison to an automatic method” PloS one, vol. 11, no. 8, p.e 0161286, 2016.

[2] M. Veta, P. J. van Diest, M. Jiwa, S. Al-Janabi, and J. P. Pluim, “Mitosis counting in breast cancer