CANCER PREDICTION SYSTEM USING PYTHON WITH DJANGO AND MACHINE LEARNING

Aniket Deshkar, Arvind Murodiya, Samrat Thakur, Mohammad Zafir Ul Hassan, Mr. Onkar G. Hastak

Department of Electronics and Telecommunication Engineering,

Priyadarshini College of Engineering, Hingna Road, Nagpur
Rashtrasant Tukdoji Maharaj Nagpur University, Maharashtra, India

I. ABSTRACT

Cancer is a disease in which certain cells in the body grow uncontrollably and spread to other parts of the body. Cancer can occur in almost any trillion cell parts of the human body. Normally, human cells will grow and multiply when needed by the body (through a process called cell division) to form new cells. As cells age or turn out to be damaged, they die, and new cells update them. Sometimes this orderly process will be disrupted, and abnormal or damaged cells will grow and multiply when they shouldn't. These cells can form tumours and are a tissue mass. Tumours can be malignant or non-malignant. Cancer can also be called malignant tumour. Many cancers form solid tumours, blood cancers such as leukaemia, but this is not the case. After removal, benign tumours usually do not grow, while cancerous tumours sometimes grow. In this project, we will build analgorithm to coach cancer histology image dataset. using Keras, we'll outline a CNN (Convolutional Neural Network), and train it on our images. We will then derive a confusion matrix to investigate the performance of the model and histology is that the study of the microscopic structure of tissues.

II. INDEX TERMS

Cancer, Machine Learning, K-Nearest Neighbor

TensorFlow, Keras, Django, NumPy

III. INTRODUCTION

Cancer is the most familiar cancer in Human of age amongst 41 to 60 and 60+, touching about 10 percent of all Human. In contemporary times, the rate keeps growing and data show that the survival rate is 88 percent after five years from diagnosis and 80 percent after 10 years from diagnosis. Early predictions of Cancer so far have made tons of improvement, death rate of Cancer by 39 percent, starting from 1989. Due to mutable nature of Cancers symptoms, patients are frequently lay open to a bombardment of assessments, comprising but not limited to mammography, ultrasound and surgery, to check their probabilities of being diagnosed with Cancer. Surgery is the most allusive among these events, which consist of intellection of sample cells or tissues for assessment. Numerical topographies, such as radius, texture, perimeter and area, can be distinguished from microscopic images. Data, later on, conquered from FNA are studied in grouping with different imaging data to prophesy probability of the patient having spiteful Cancer tumour. A computerized system here would be colossally profitable in this situation. It will possibly speed up the process and enhance the meticulousness of the doctor's predictions. In addition, if supported by plethora dataset and the computerized system dependably carry out well, it will conceivably disregard the necessities for patients to go through copious of other tests, such as mammography, ultrasound, and MRI, which focus patients to major extent of soreness and radiation. In all, an early calculation remains is one of the vigorous features in the follow-up process. Data extracting techniques or sorting can help to lessen the number of false positive and false negative assessments. As a result, a new method like data discovery in databases has become a preferential implement for medical assistant.

Given the significance of customized medicine and the escalating idea on the application of ML techniques, we here present ananalysis ofreadingsthatmakeuseofthesemethodologieson the issues ofthecancerprediction.Inthesereadingspredictivetopographies are measured which may be autonomous of a convinced treatment or arecombinedinordertoattend dealingforcancerpatients,respectively.Inaddition,wediscussthetypesofMLtechniquesbeingused,the varietiesofdatatheyassimilate,theoverallenactmentofeachproposed outline. while we also discuss their advantages and disadvantages.

Different systems that could enable the early cancer analysis. Specifically, these readings refer to methods interconnected to the profiling of interspersing miRNAs that have been confirmed a promising class for cancer recognition and identification. However, these methodologies suffer from low understanding with reference to their use in showing at early phase and their difficulty to distinguish benevolent from malevolent tumors. Numerous facets regarding the prediction of cancer result based on gene expression emblems are deliberated in. These readings list the possible as well as the confines of microarrays for the prediction of cancer cause. Even though gene signatures could significantly make progress our capacity for prediction in cancer patients, poor improvement has been made for their use in the treatment center. However, before gene expression profiling can be used in clinical practice,
Studies with larger data sections and more adequate demonstrationare desirable.

ML,adivisionofArtificialIntelligence,recountsthedelinquentoflearning fromdatasamplestothecommontheoryofimplication.Each learningprocessconsistsoftwosegments:(I)assessmentofunidentifieddependenciesinasystemfromagivendatasetand(ii)useofassessed dependenciestoforewarningnewoutputsofthesystem. MLhasalsobeen confirmedaremarkableareainbiomedicallearningwith manysolicitations, where an adequate summary is conquered by using diverse techniques and algorithms.

IV. LITERATURE SURVEY

Many works are submitted that tried to diagnose carcinoma mistreatment machine learning algorithms. For instance, Sun at al. in year 2005, planned examination feature selection ways for a unified detection of cancers. Another approach, introduced by Malek at al. in year 2009, planned a way mistreatment ripple and proposed a style of machine-controlled detection, segmentation, and classification of carcinoma nuclei employing a symbolic logic for feature demand and classification respectively. Zheg at al. in year 2014 combined support vector machine (SVM) and K-means algorithmic program for cancer diagnosis. Aličković and Subasi in year 2017 applied a genetic algorithmic program for feature extraction and rotation for classification. Another approach is conducted by Bannaie in year 2018 supported the dynamic contrast-enhanced resonance imaging (DCE-MRI) technique to realize output of interest. There are many different works performed based on agglomeration and classification. Alireza Osarech, Bita Shadgar achieved 98.80% and 96.63% curacies upon mistreatment SVM classification technique on 2 completely different benchmark datasets for carcinoma. Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza applied KNN, SVM, Gaussian Naïve Bayes, and logistical Regression techniques programmed in MATLAB to diagnose and predict repeat of cancer. The classification techniques were applied on 2 datasets from UCI depository. One dataset was used for identification of diseases (WBCD), and different is employed for prediction of repeat. Vikas Chaurasia, BB shot Tiwari and Saurabh Pal build prognostic models on carcinoma and compared their accuracies mistreatment far-famed algorithms videlicet J48, Naïve Bayes, and RBF. The results indicated that Naïve Thomas Bayes expected well among them with 97.36accuracy. Haifeng Wang and Sang Won Yoon developed a robust model for carcinoma prediction by mistreatment and comparing Naive Thomas Bayes Classifier, Support Vector Machine (SVM), AdaBoost tree and Artificial Neural Networks (ANN). They enforced PCA for spatiality reduction.S. Kharya planned Artificial Neural Networks (ANN) whereas functioning on carcinoma prediction. The paper highlighted blessings of mistreatment machine learning ways like SVM, Naive Bayes, Neural network and call trees. Naresh Khuriwal and Nidhi Mishra used Wisconsin carcinoma info to figure on cancer diagnosis. supported their experiments they finished that, ANN and logistical algorithmic program worked higher and achieved an accuracy of 98.50%.

V. BLOCK DIAGRAM

Figure 1 INPUT STAGE

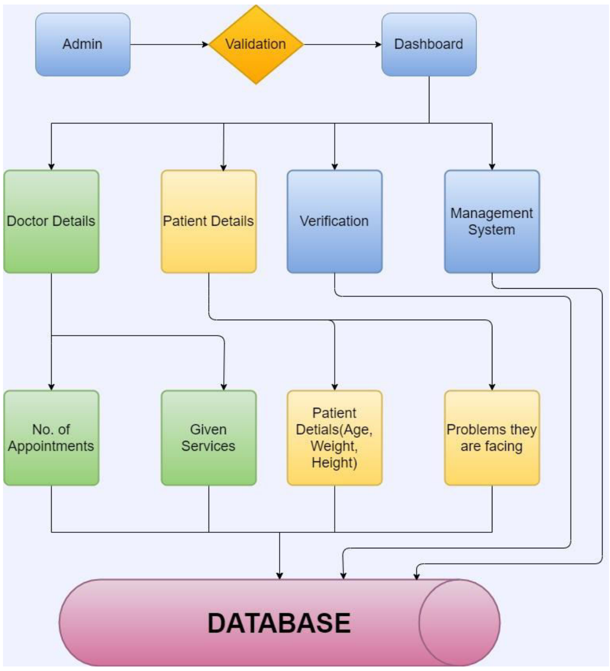


Figure 2 PROCESSING STAGE

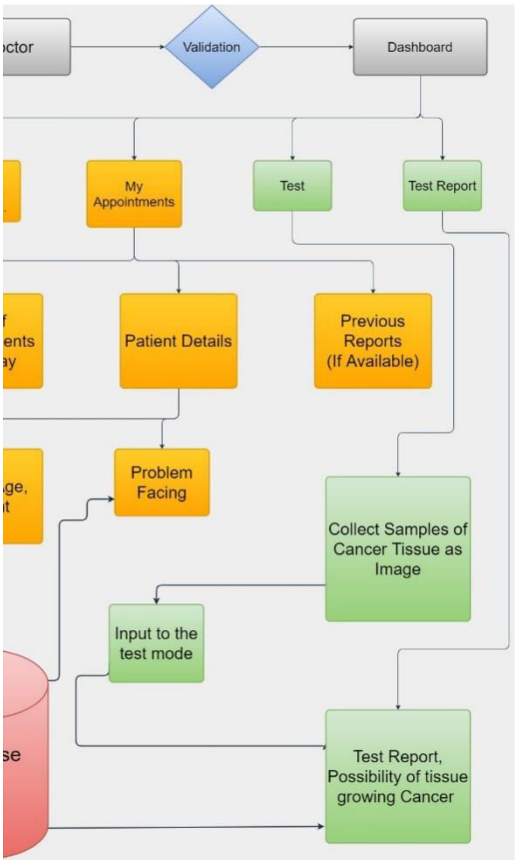
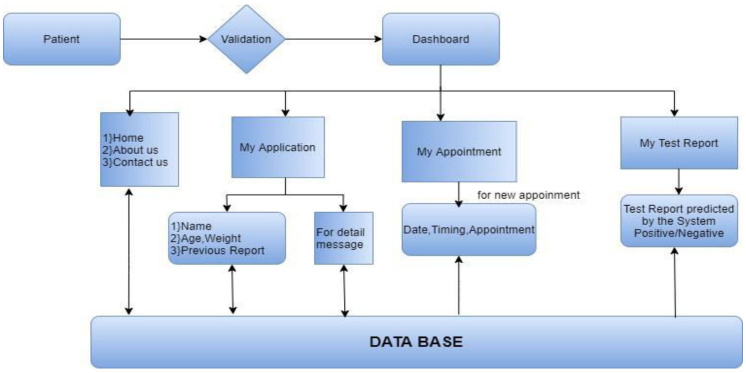

ctor
Validation
Dashboard
My Appointments
Test
Test Report
f
ents
ay
Patient Details
Previous Reports (If Available)
ge,
t
Problem Facing
Collect Samples of Cancer Tissue as Image
Input to the test mode
se
Test Report, Possibility of tissue growing Cancer

Figure 3 OUTPUT STAGE



Aasda .

VI. PROJECT WORKING

Firstly, we intend to do is to gather the information that we have an interest in assembling for pre-processing and to use classification and regression methods. The information collected is assembled in pre-processing which is called data processing technique that involves transforming information into a clear format. All the available data is commonly incomplete, inconsistent, and lacking bound to contain several errors. Data pre-processing technique could be a proved by methodology of breaking down such issues. For pre-processing we've used standardized method to pre-process the UCI dataset. This step is incredibly vital as a result of the standard and amount of information that you gather can directly confirm how much smart and accurate your prophetical model will be. In our case we tend to gather all the data related to the Cancer samples, this can be our coaching data.

In Preparation, wherever we tend to load our informative data into an appropriate place and prepare it to be used in our machine learning training. We'll place all our data together, and so disarrange the ordering.

In machine learning and statistics, feature selection, additionally referred to as variable selection, attribute selection, is the process of choice a set of relevant options to be used in model construction. Data File of cancer samples that we've got is used in Wrapper methodology for Feature Selection. The vital options found by the study are: pouch-shaped points worst, space worst, space se, Texture worst, Texture mean, Smoothness worst, Smoothness Radius mean, Symmetry mean.

Feature Projection is transformation of high-dimensional area of information to a lower dimensional space (containing few attributes), each linear and nonlinear reduction techniques is employed in accordance in related relationships among the options within the dataset.

Feature Scaling Most of the times, your dataset can contain options extremely varied in magnitudes, units and range. However since, most of the machine learning algorithms use geometer distance between two different points in their computations. We need to bring all options to constant level of magnitudes. This will be achieved by scaling.

Model choice Supervised learning is that the methodology within which the machine is trained on the informative cancer samples which the input and output are well labelled. The model can learn on the coaching knowledge and may method the longer term data to predict outcome. They are sorted to Regression and Classification techniques. Now In unattended learning algorithmic program the machine is trained from the information that isn't tagged or classified creating the algorithmic program to figure while is correct or incorrect instructions. In our dataset we've got the result variable or variable quantity. Thus Classification of algorithmic program of supervised learning is applied on it. We've got chosen three differing kinds of classification algorithms in Machine Learning.

Prediction Machine learning is victimization knowledge to answer questions. Thus Prediction, or inference, is that the step wherever we tend to get to answer some questions. This is often the purpose of all this work, wherever the worth of machine learning is real and accurate.

VII. TECHNOLOGIES USED

Python:

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation.

Django:

Django is a Python-based free and open-source web framework that follows the model–template–views architectural pattern.

SQLite3:

SQLite is a relational database management system contained in a C library. In contrast to many other database management systems, SQLite is not a client–server database engine. Rather, it is embedded into the end program.

JavaScript:

JavaScript is high-level, often just-in-time compiled, and multi-paradigm. It has curly-bracket syntax, dynamic typing, prototype-based object-orientation, and first-class functions

HTML/CSS3:

The Hypertext Markup Language, or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets and scripting languages such as JavaScript.

TensorFlow:

an open-source software library that uses mathematical algorithms that can handle tensor operations. The library expresses output in graphs and n-dimensional matrix. It has modularity attributes, which makes it flexible and is easy to use for training architecture. TensorFlow can process and train multiple networks. Therefore, it is useful when working with a larger system

Keras:

a neural network library in Python that uses TensorFlow in the backend infrastructure to compile models and graphs for machine learning. It can be implemented in almost every neural network model and be processed on both CPU and GPU with high speed. Keras is commonly used and is easy to implement when working with images or text.

Scikit-learn:

a Python library that can be used when implementing algorithm(s) in difficult model training. It contains numerous functions, such as classification and model selection. The library has features like cross-validation that calculates the accuracy of the model.

NumPy:

A machine learning library that works along with other libraries to perform array operations. The library is easy to use for complex mathematical implementations. These features are applied when working with expressing binary in an array of n-dimensions, images, or sound-waves

Panda:

A library in Python that has multiple different features for analysis of data structure. The library has built-in functionalities such as translation of operation and data manipulation, which provides flexibility with high functionality.

VIII. RESULT

The results of cross validation of each model are compared against coaching and testing set. Taking confusion matrix into thought and analysing the accuracies, it's ascertained that although, SVM linear model is slightly unbalanced with cross validation 97.19%, coaching set 98.83% and testing 96.50%, however this may be generalized by ever-changing and modifying the training and testing set. the number of observations in coaching and testing set additionally has significant impact on the accuracy of data. Thus, with performance metric of 97.19%, SVM linear is relatively a lot of accurate than linear Regression, KNN, and Decision Tree in detection of cancer.

IX. FUTURE SCOPE

In this thesis, genomic sequencing and image process strategies were enforced to sight and predict cancer of the blood in knowledge samples. any add this space will be exploitation completely different neural network architecture and solely exploitation one dataset. this might be fascinating to look at and compare that networks formula would have higher performances. alternative sorts of validations splits may even be used to check out and analyze the impact it may wear the models results. Furthermore, making a way to change the pre-processing step for the genomic sequence may be one thing to figure on, to reduce the manual portion in this phase. it'd contribute to the chance of accelerating the samples to the dataset and check the accuracy distinction between the methods.

X. CONCLUSION

This study makes an attempt to analyse numerous supervised machine learning algorithms and choose the most correct model in detection of breast cancer. The work targeted in advancement of predictive models with the assistance of python to realize higher accuracy in predicting correct outcomes. The analysis of result signifies that, integration of data, feature scaling at the side of different classification technique and analysis gives markedly flourishing tool in prediction. It has additionally discovered that the model misdiagnosed few patients with cancer once they weren't having cancer and vice versa. Although, the model is correct however once coping with lives of people, any analysis in building the foremost correct and precise model should be allotted for higher performance of classification techniques and acquire the accuracy as on the point of 100 percent as possible. Thus, the standardization of every of the models is important with the building of a lot of reliable model.

ACKNOWLEDGEMENT

This project is being implemented by students namely Aniket Deshkar, Samrat Thakur, Zafir Ul Hassan, Arvind Murodiya under the guidance of Prof. Mr. Onkar G. Hastak at Department of Electronics and Telecommunication of Priyadarshini College of Engineering,Nagpur, under R.T.M. Nagpur university, Nagpur, Maharashtra, India

REFERENCES

1. Y. Sun, C. F. Babbs, and E. J. Delp, "A comparison of feature selection methods for the etection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm," in Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 6532–6535, Shanghai, China, September 2005.

2. J. Malek, A. Sebri, S. Mabrouk, K. Torki, and R. Tourki, "Automated breast cancer diagnosis based on GVF-snake segmentation, wavelet features extraction and fuzzy classification," Journal of Signal Processing Systems, vol. 55, no. 1–3, pp. 49–66, 2009.

3. B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Systems with Applications, vol. 41, no. 4, pp. 1476–1482, 2014.

4. E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," Neural Computing and Applications, vol. 28, no. 4, pp. 753–763, 2017. 19. M. Banaie, H. Soltanian-Zadeh, H.-R. Saligheh-Rad, and M. Gity, "Spatiotemporal features of DCE-MRI for breast cancer diagnosis," Computer Methods and Programs in Biomedicine, vol. 155, pp. 153–164, 2018.

5. M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 36, no. 2, pp. 3240–3247, 2009.

6. Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. Journal of Biomedical Informatics, 69, 218-229. Doi: 10.1016/j.jbi.2017.04.001

7. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

8. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101