INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1964)
Available online at: https://www.ijariit.com

Disease detection using Machine Learning

Jasmeet Singh Khokhar
jasmeetsinghkhokhar@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Ankush Aglawe
ankushaglawe123@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Akash Pandey
ap214708@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Narayana Naidu
narayananaidu541@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Roshan Mankar
roshanmankar002@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

ABSTRACT

Cancer is a lethal disease produced by the aggregation of hereditary disorders and a variety of pathological alterations. Cancerous cells are life-threatening abnormal regions that can grow in any portion of the human body. Cancer is also known as a tumour that must be diagnosed swiftly and accurately in the early stages in order to determine what treatmentoptions are available. Despite the fact that each modality has its own set of problems, such as a difficult history, incorrect diagnoses, and therapy, which are all major causes of death. The goal of the study is to examine, review, evaluate, and discuss current breakthroughs in human body cancer detection utilising machine learning approaches for breast, brain, lung, liver, and skin cancers, as well as leukaemia. The study shows how machine learning with supervised, unsupervised, and deep learning techniques can help in cancer diagnosis and cure. Several state-of-the-art approaches are grouped together, and findings fromaccuracy, sensitivity, specificity, and false-positive metrics are compared on benchmark datasets. Finally, potential future workis indicated by highlighting obstacles.

Keywords: *Activation Function, Convolution Neural Network, Datasets, Data Preprocessing, Image Classification, Sigmoid Function, X-Rays.*

1. INTRODUCTION

Healthcare is one of the most pressing issues in human civilizations, as it directly affects residents' quality of life.The healthcare industry, on the other hand, is exceedinglydiversified, dispersed, and fragmented. Clinically, provid-ing optimal patient treatment necessitates access to rele- vant patient information, which is rarely available. Health-care is one of the most pressing issues in human civiliza- tions, as it directly affects residents' quality of life. The healthcare industry, on the other hand, is exceedingly di- versified, dispersed, and fragmented. Clinically, provid- ing optimal patient care necessitates access to relevant pa-tient data, which is rarely available when and where it is needed. Additionally, the wide variation in test-ordering for diagnostic purposes suggests the requirement of a suf-ficient and appropriate test set. Smellie et al. extended this idea by claiming that the substantial discrepanciesin general practise pathology seeking are primarily dueto individual diversity in clinical practise and, as a re- sult, may be changed by doctors making more consistent and well-informed decisions. As a result, medical data is frequently made up of a large number of heterogeneous variables gathered from various sources, such as demo- graphics, disease history, medication, allergies, biomark-ers, medical images, or genetic markers, each of which provides a unique perspective on a patient's condition. Furthermore, the statistical features of the aforementioned sources differ intrinsically. When academics and practitioners analyse such data, they face two issues: the curse of dimensionality (the feature space grows exponentially in terms of the number of dimensions and samples) and vari-ability in feature sources and statistical attributes. These factors provoke delays and inaccuracy in the disease de- tection and, consequently, patients could not receive the appropriate care. Thus, there is a clear need for an ef- fective and robust methodology that allows for early dis- ease detection and it can be used by doctors as help for decision-making. Therefore, medical, computational, and statistical fields are facing the challenge of exploring newtechniques for modeling the prognosis and diagnosis of diseases, since traditional paradigms fail in the treatment of all this information. This necessity is closely linked to developments in other fields such as Big Data (BD), DataMining (DM), and Artificial Intelligence (AI). The scienceof data management and analysis is advancing to convert this massive repository of information and knowledge that supports them in reaching their goals, as the amount of medical data being digitally collected and stored is mas- sive and expanding rapidly. BD is the name given to this fast growing technology.

2. SCOPE AND OBJECTIVE

Different types of Disease detection and classification us-ing machine assistance have opened up a new research area for early detection of disease, which has shown the ability to reduce

© 2021, www.IJARIIT.com All Rights Reserved

manual system impairments. This survey presents several sections on state of art techniques, anal- ysis, and comparisons on benchmark datasets for breast cancer, lung cancer, skin lesion detection, and malaria detection respectively from F-measure, sensitivity, speci- ficity, accuracy, precision points of view. The pictorial depiction of this study is presented.

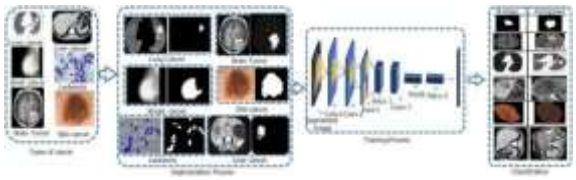

Fig. 1: Machine assisted system for cancer detection.

3. BENCHMARK AND DATASETS

Lung Cancer

Nodules in Chest X-rays (LIDC-IDRI) Kaggle

- A curated chest X-ray dataset with nodule anno- tations done by hand. Four radiologists who had access to the patient's CT image marked nodules. Even though the nodule itself would not be seen onX-ray, each radiologist was required to write an an-notation (but CT scan indicated otherwise).
- 3.8 GB of dataset 463 images files: Attackers can intercept and add/remove medical evidencein medical imagery with high realism using deep learning.We show medical deepfakes in this dataset: 3D CT scans of human lungs, some of which have been altered with genuine cancer removed and fake cancer injected. The goal of this dataset is to identify where medical scans havebeen altered with and discriminate between true and fraud-ulent malignancies. Three professional radiologists and a cutting-edge AI analysed this dataset and were unable to consistently distinguish between real and phoney tumours, implying that the fake tumours appear lifelike and detec- tion is tough[2].

Skin Lesion Detection

SIIM-ISIC-melanoma-classification (skin lesion images)

- The International Skin Imaging Collaboration (ISIC) created the dataset, which includes images from the Hospital Clinic de Barcelona, the Medi- cal University of Vienna, Memorial Sloan Ketter- ing Cancer Center, Melanoma Institute Australia, The University of Queensland, and the University of Athens Medical School[4].
- Image file of dataset 108.19 GB in jpeg format
- Train - the training set
- Test - the test set
- 88.3k files and 15 columns

Malaria Detection

Cell-images-for-detecting-malaria: This Dataset is taken from the official NIH Website anduploaded here, so anybody trying to start working with this dataset can get started immediately, as downloadingthe dataset from NIH website is quite slow[3]. The dataset contains 2 folders

- Infected
- Uninfected

And a total of 27,558 images. 335 MB size having 27.6k files

Breast Cancer Detection Breast Histopathology image

- The original files are located at gleason website, 1.4GB size having 278k files
- Breast cancer is the most frequent cancer in women,and the most prevalent type of breast cancer is inva-sive ductal carcinoma (IDC). Automated approaches can be utilised to save time and reduce error for detecting and categorising breast cancer subtypes, which is a crucial clinical activity[1].

4. IMPLEMENTATION DETAILS

When it comes to Machine Learning, Artificial Neural Networks perform well. Artificial Neural Networks are used in various classification tasks like image, audio, 10 words. Neural Networks come in a variety of shapes and sizes, and they're utilised for a variety of applications. Forexample, to forecast the sequence of words, we use Re- current Neural Networks, more precisely an LSTM, and for image classification, we use Convolution Neural Net- works. In this paper, we are going to build a basic build- ing block for CNN. Before we get into the Convolution Neural Network, let's go over some basic neural network ideas. There are three sorts of layers in a standardNeu-ral Network: Input Layers: It's the layer in which we give input to our model. The entire number of character- istics in our data is equal to the number of neurons in thislayer (number of pixels in case of an image). Hidden Layer: The input from the Input layer is then fed into the hidden layer. Depending on our model and data size, therecould be a lot of hidden layers. The number of neurons ineach hidden layer can vary, though they are usually more than the number of characteristics. The output of each layer is calculated by matrix multiplication of the preced-ing layer's output with that layer's learnable weights, fol- lowed by the addition of learnable biases, and finally the activation function, which makes the network nonlinear. Output Layer: The hidden layer's output is then passedinto a logistic function like sigmoid or softmax, which converts each class's output into probability scores. Af- ter that, the data is input into the model, and each layer's output is received. This step is called feed-forward, we then calculate the error using an error function, some com-mon error functions are cross-entropy, square loss error, etc. After that, we calculate the derivatives to backpropa- gate into the model. This step is called Backpropagation which is used to minimize the loss.

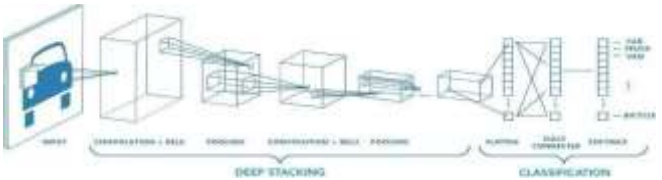

Fig. 2: Convolution Neural Network Implementation.

Types of layers

Let's take an example by running a covnets on an image of dimension 32 x 32 x 3.

1. *Input Layer:* This layer holds the raw input of an image with width 32, height 32 and depth 3.
2. *Convolution Layer:* The output volume is computed by computing the dot product of all filters and the image patch in this layer. Suppose we use a total 12filters for this layer we'll get an output volume of dimension 32 x 32 x 12.
3. *Activation Function Layer:* This layer will apply el-ement wise activation function to the output of the convolution layer. Some common activation func- tions are RELU: max(0, x), Sigmoid: 1/(1+e-x), Tanh, Leaky RELU, etc. Because the volume stays unchanged, the output volume will be 32 x 32 x 12.
4. Pool Layer: This layer is added into the covnets on aregular basis, and its major goal is to lower the vol-ume size, which speeds up calculation, saves memory, and prevents overfitting. Max pooling and aver-age pooling are two typical types of pooling layers.If we utilise a maximum pool with 2 x 2 filters and stride 2, the final volume will be 16x16x12.
5. *Fully-Connected Layer:* This layer is a typical neu-ral network layer that takes input from the previouslayer, computes class scores, and outputs a 1-D ar- ray with the same number of classes as the previouslayer.

© 2021, www.IJARIIT.com All Rights Reserved

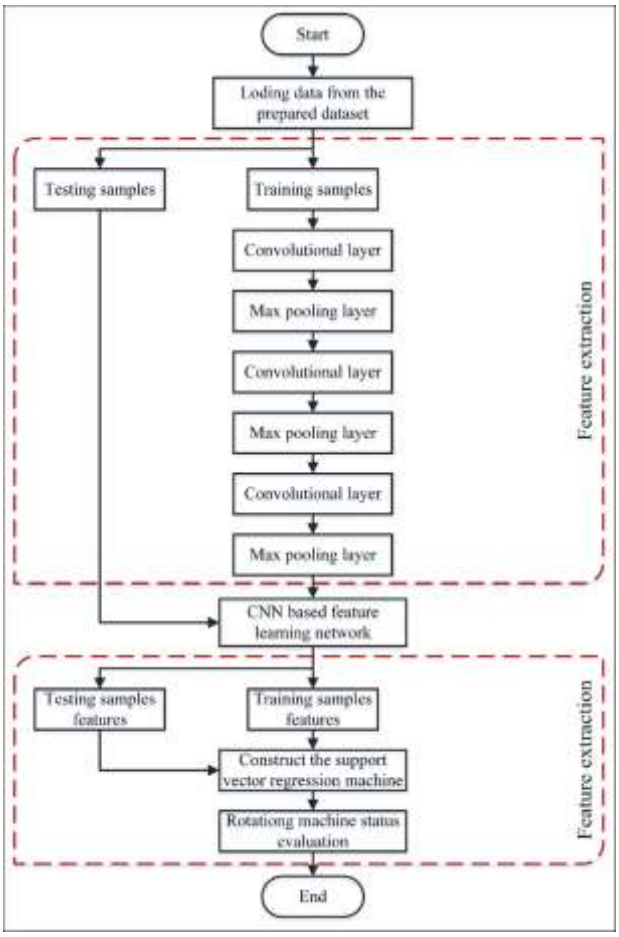

Fig. 3: Flow chart of CNN.

5. CONCLUSION

The past few decades have witnessed a revolution in the detection and cure of cancer using machine assistance. Ac- cordingly, this paper has presented a systematic review ofcurrent techniques in the diagnosis and cure of several can- cers affecting the human body badly. The purpose of this article is to examine, assess, and categorize the approachesused to treat various types of cancer, as well as to iden- tify any existing limitations. The review has presented different types of cancers lung cancer, breast cancer, and skin cancer. Additionally, this study has presented the significant stages of automated cancer diagnosis such as image pre-processing and classification using benchmark datasets. The fundamental goal of this study is to provide new researchers with a thorough background in order for them to begin their research in this subject.

Finally, a thorough evaluation of current state-of-the-art machine-assisted cancer detection strategies, as well as their benefits and drawbacks. However, each cancer cat- egory's accuracy is still far from mature. The majority of the researchers didn't use benchmark datasets or only usedtiny samples to test their proposed strategies. The present state of the art methodologies are contrasted on benchmark datasets for this purpose, and the shortcomings of present strategies are emphasized.

The main challenges in the cancer detection and cure process redesign the research pipeline, understanding the cancer growth phenomena, developing preclinical models, handling complex cancers precisely, early treatment, inno- vative methods of designing and delivering clinical trials,and enhancing accuracy that will be useful for the physi- cians as a second and early opinion.

6. REFERENCES

[1] Meriem Amrane, Saliha Oukid, Ikram Gagaoua, and Tolga Ensari. Breast cancer classification using ma- chine learning. In *2018 Electric Electronics, Com- puter Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4. IEEE, 2018.

[2] Savannah L Bergquist, Gabriel A Brooks, Nancy L Keating, Mary Beth Landrum, and Sherri Rose. Clas- sifying lung cancer severity with ensemble machine learning in health care claims data. In *Machine Learn- ing for Healthcare Conference*, pages 25–38. PMLR,2017.

[3] Zhaohui Liang, Andrew Powell, Ilker Ersoy, Mahdieh Poostchi, Kamolrat Silamut, Kannappan Palaniap- pan, Peng Guo, Md Amir Hossain, Antani Sameer, Richard James Maude, et al. Cnn-based image anal- ysis for malaria diagnosis. In *2016 IEEE Interna- tional Conference on Bioinformatics and Biomedicine (BIBM)*, pages 493–496. IEEE, 2016.

[4] Ilker Ali OZKAN and Murat KOKLU. Skin lesion classification using machine learning algorithms. *In- ternational Journal of Intelligent Systems and Appli- cations in Engineering*, 5(4):285–289, 2017.

© 2021, www.IJARIIT.com All Rights Reserved