CNN APPROACH FOR CANCER PROGNOSIS USING MACHINE LEARNING

Mr. Dr. B. Sunil Kumar Ph.D	D. Arpitha Singh	P. Kruthik Shirin	B. S. S Khyathi	SK.Jasmeen
Computer Science and Engineering Narayana Engineering College(JNTUA) Nellore,India	*Computer Science and Engineering Narayana Engineering College(JNTUA)* Nellore,India	*Computer Science and Engineering Narayana Engineering College(JNTUA)* Nellore,India	*Computer Science and Engineering Narayana Engineering College(JNTUA)* Nellore,India	*Computer Science and Engineering Narayana Engineering College(JNTUA)* Nellore,India
Sunny.bs2003@gmail.com	dheerajarpithasingh@g mail.com	shirin.paduchuri@gmail.com	Khyathisarma13@gmail.com	Shaikjasmeen05@gmail.com

Abstract: **Among different illnesses, malignant growth has become a major danger to people internationally. Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. Machine learning is frequently used in cancer diagnosis and detection. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning methods. These techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. Here we are using convolutional neural networks to predict the cancer qualities. It is an algorithm to analyzing visual imagery. In this project, we use several Convolutional Neural Network models that take unstructured gene expression inputs to classify tumor and non-tumor samples into their designated cancer types are as normal.**

Keywords **-- Cancer, Machine learning, CNN**

1.INTRODUCTION:

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. Machine Learning is one of the most exciting technologies that one would have ever come across. As it is evident from the name,it gives the computer that makes it more similar to humans: *The ability to learn itself from experience.*

Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

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 techniques and algorithms. There are two main common types of ML methods known as (i) supervised learning and (ii) unsupervised learning. In supervised learning a labelled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labelled examples are provided and there is no notion of the output during the learning process.

Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. A tumor is an abnormal mass of cells. Tumor is of two types known as (i) Malignant and (ii) Benign. A Malignant tumor means it is made of cancer cells, and it can invade nearby tissues. In contrast a benign tumor is a mass of cells that lacks the ability to either invade neighboring tissue.

A convolutional Neural Network is a class of deep neural networks, most commonly applied to analysing visual imagery. They are also known as shift invariant or space invariant artificial neural networks, based on their shared weights architecture and translation invariance characteristics.

The most common use of CNN is image classification, for example identifying satellite images that contain roads or classifying hand written letters and digits. There are other quite mainstream tasks such as image segmentation and signal processing, for which CNN's perform well. CNN's have been used for understanding in Natural Language Processing (NLP) and Speech recognition, although often for NLP Recurrent Neural Nets are used. A CNN can also be implemented as a U-Net architecture, which are essentially two almost mirrored CNN's resulting in a CNN whose architecture can be presented in a U-shape. U-nets are used where the output needs to be of similar size to the input such as segmentation and image improvement.

II. RELATED WORK

Cancer is the second leading cause of death worldwide, an average of one in six deaths is due to cancer. Considerable research efforts have been devoted to cancer diagnosis and treatment techniques to lessen its impact on human health. Cancer prediction's major focus is on cancer susceptibility, recurrence, and prognosis, while the aim of cancer detection is the classification of tumor types and identification of markers for each cancer such that we can build a learning machine to identify specific metastatic tumor type or detect cancer at their earlier stage.

Cancer diagnosis

- **Physical exam.** Your doctor may feel areas of your body for lumps that may indicate a tumor. During a physical exam, he or she may look for abnormalities, such as changes in skin color or enlargement of an organ, that may indicate the presence of cancer.
- **Laboratory tests.** Laboratory tests, such as urine and blood tests, may help your doctor identify abnormalities that can be caused by cancer. For instance, in people with leukemia, a common blood test called complete blood count may reveal an unusual number or type of white blood cells.
- **Imaging tests.** Imaging tests allow your doctor to examine your bones and internal organs in a noninvasive way. Imaging tests used in diagnosing cancer may include a computerized tomography (CT) scan, bone scan, magnetic resonance imaging (MRI), positron emission tomography (PET) scan, ultrasound and X-ray, among others.
- **Biopsy.** During a biopsy, your doctor collects a sample of cells for testing in the laboratory. There are several ways of collecting a sample. Which biopsy procedure is right for you depends on your type of cancer and its location.In most cases, a biopsy is the only way to definitively diagnose cancer.

  In the laboratory, doctors look at cell samples under the microscope. Normal cells look uniform, with similar sizes and orderly organization. Cancer cells look less orderly, with varying sizes and without apparent organization.

  There is a machine learning algorithm called SVM (support vector machine) which is used to predict the cancer. But the main disadvantage of SVM algorithm is not suitable for large data sets.

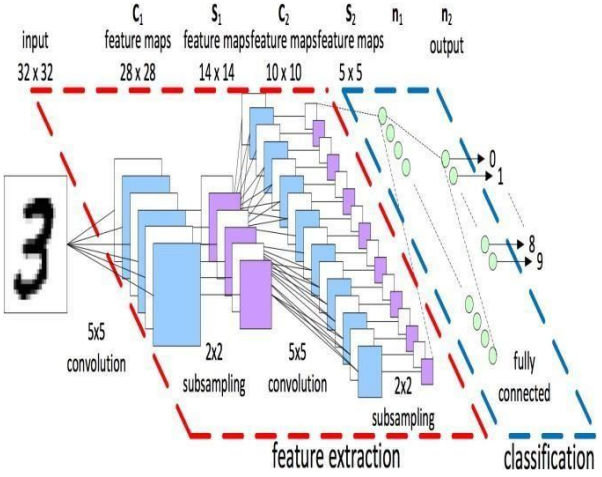
  Disadvantages:

  - False cancer detection is also present in modern diagnosis
  - Linked to a natural anxiety of specialists to avoid overlooking cancer at earlier stages.

III. PROPOSED WORK

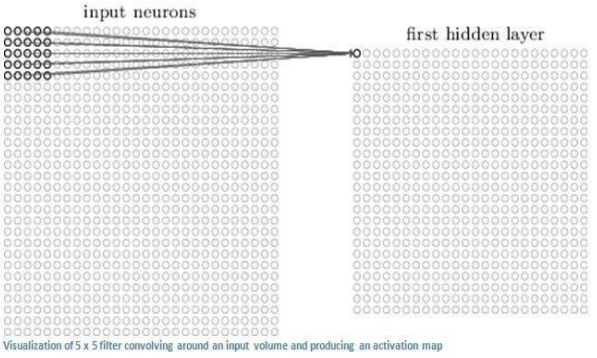
CNN (convolutional neural networks) have a different architecture than regular Neural Networks. Regular Neural Networks transform an input by putting it through a series of hidden layers. Every layer is made up of a set of neurons, where each layer is fully connected to all neurons in the layer before. Different CNN models were proposed for cancer type prediction. Each model aims to address a specific aspect of cancer cells. Conventionally, the first ConvLayer is responsible for capturing the Low-Level features such as edges, color, gradient orientation, etc. With added layers, the architecture adapts to the High-Level features as well, giving us a network which has the wholesome understanding of images in the data-set, similar to how we would.

Convolution in CNN is performed on an input image using a filter or a kernel. To understand filtering and convolution you will have to scan the screen starting from top left to right and moving down a bit after covering the width of the screen and repeating the same process until you are done scanning the whole screen



When a computer sees an image, it will see an array of pixel values. Depending on the resolution and size of the image, it will see a 32 x 32 x 3 array of numbers. Just to drive home the point, let's say we have a colour image in JPG form and its size is 480 x 480. The representative array will be 480 x 480 x 3. Each of these numbers is given a value from 0 to 255 which describes the pixel intensity at that point. These numbers, while meaningless to us when we perform image classification, are the only inputs available to the computer.The idea is that you give the computer this array of numbers and it will output numbers that describe the probability of the image being a certain class.

The first layer in a CNN is always a Convolutional Layer. First thing to make sure you remember is what the input to this convlayer is. Like we mentioned before, the input is a 32 x 32 x 3 array of pixel values. Now, the best way to explain a conv layer is to imagine a flashlight that is shining over the top left of the image. Let's say that the light this flashlight shines covers a 5 x 5 area. As the filter is sliding, or convolving, around the input image, it is multiplying the values in the filter with the original pixel values of the image.



Visualization of 5 x 5 filter convolving around an input volume and producing an activation map

CNNs specifically are inspired by the biological visual cortex. The cortex has small regions of cells that are sensitive to the specific areas of the visual field. This idea was expanded by a captivating experiment done by Hubel and Wiesel in 1962. In this experiment, the researchers showed that some individual neurons in the brain activated or fired only in the presence of edges of a particular orientation like vertical or horizontal edges. For example, some neurons fired when exposed to vertical sides and some when shown a horizontal edge. Hubel and Wiesel found that all of these neurons were well ordered in a columnar fashion and that together they were able to produce visual perception. This idea of specialized components inside of a system having specific tasks is one that machines use as well and one that you can also find back in CNNs.

Convolution has the nice property of being translational invariant. Intuitively, this means that each convolution filter represents a feature of interest and the Convolutional Neural

Network algorithm learns features comprise the resulting reference.

We have 4 steps for convolution:

- Line up the feature and the image
- Multiply each image pixel by corresponding feature pixel
- Add the values and find the sum
- Divide the sum by the total number of pixels in the feature

After a convolution layer once you get the feature maps, it is common to add a pooling or a sub-sampling layer in CNN layers. Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to decrease the computational power required to process the data through dimensionality reduction. Furthermore, it is useful for extracting dominant features which are rotational and positional invariant, thus maintaining the process of effectively training of the model. Pooling shortens the training time and controls over-fitting.

There are two types of Pooling:

Max Pooling:

Max Pooling and Average Pooling. Max Pooling returns the maximum value from the portion of the image covered by the Kernel.

Max Pooling also performs as a Noise Suppressant. It discards the noisy activation altogether and also performs de-noising along with dimensionality reduction.

Average Pooling:

Average Pooling returns the average of all the values from the portion of the image covered by the Kernel.

Average Pooling simply performs dimensionality reduction as a noise suppressing mechanism. Hence, we can say that Max Pooling performs a lot better than Average Pooling.

The Convolutional Layer and the Pooling Layer, together form the i-th layer of a Convolutional Neural Network. Depending on the complexities in the images, the number of such layers may be increased for capturing low-levels details even further, but at the cost of more computational power.

After going through the above process, we have successfully enabled the model to understand the features. Moving on, we are going to flatten the final output and feed it to a regular Neural Network for classification purposes.

IV. CONCLUSION

In this project, CNN architecture that take high dimension cancer scanned image inputs and perform cancer type prediction while considering their tissue of origin. Our model achieved an equivalent 95.7% prediction accuracy -comparing to earlier published studies, however with a drastically simplified CNN construction and with a significant reduction from tissue of origin. This allows us to perform a normal interpretation of our CNN model to elucidate cancer markers for each cancer type, with hope in future refinement that will lead to markers for earlier cancer detection.

V. REFERENCES

1. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2018. CA Cancer J Clin 2018, 68(1):7-30.
2. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A et al: Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018, 359(6378):926-930.
3. Haendel MA, Chute CG, Robinson PN: Classification, Ontology, and Precision Medicine. N Engl J Med 2018, 379(15):1452-1462.
4. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, Anagnostou V, Fiksel J, Cristiano S, Papp E et al: Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med 2017, 9(403):eaan2415.
5. Schiffman JD, Fisher PG, Gibbs P: Early detection of cancer: past, present, and future. Am Soc Clin Oncol Educ Book 2015:57-65.
6. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 2015, 521(7553):436-444.

7. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM: Toward a Shared Vision for Cancer Genomic Data. N Engl J Med 2016, 375(12):1109-1112.
8. Ahn T, Goo T, Lee C-h, Kim S, Han K, Park S, Park T: Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 2018: IEEE; 2018: 1748-1752.
9. Lyu B, Haque A: Deep learning based tumor type classification using gene expression data. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics: 2018: ACM; 2018: 89-96.
10. Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, Li L: A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC Genomics 2017, 18(1):508.