International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

Comparative Analysis of Early Stage Cancer Detection Methods in Machine Learning

Seeyog Kapadne[1], Arsh Patne[2], Jahan Chaware[3] and Prof. Priyanka Shahane[4]

[1]*Student, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, Maharashtra, India*
[2]*Student, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, Maharashtra, India*
[3]*Student, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, Maharashtra, India*
[4]*Assistant Professor, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, Maharashtra, India*

Abstract - *Cancer is a collection of diseases, which is driven by changes in cells of the body by increasing normal growth and control. Its prevalence is increasing yearly and is advancing along with it to counter the occurrences and provide solutions.*
The early stages of cancer detection are required to provide proper treatment to the patient and reduce the risk of death due to cancer as detection of these cancer cells at later stages leads to more suffering and increases the chances of death.
This research aims to study various techniques for detecting cancer in its early stages.

Key Words: **Convolutional Neural Network, Random Forest, Machine Learning, Linear Regression**

1. INTRODUCTION

Cancer is one of the major diseases which needs to be taken care of in its early stages; otherwise, the excess cancer cells cause the most damage to the body and weaken the person. It is a priority to detect these cancer cells at an early stage to cure them simple and cause no harm to the person's life by those cells. If we find cancer cells before proceeding to further stages, then we will be capable to retain many lives. Many people cannot afford to spend money to cure this cancer or to test it, so our main aim is to take this test for as low a cost as possible so that everyone will be able to afford it and be able to cure it at an early stage with no harm to their lives. To help us with all this, we need to make a system that would help us detect this cell and give output accordingly, so machine learning can be an option here to guess these cells and provide an accurate yield. This survey paper presents all types of algorithms as supervised, semi-supervised, and unsupervised machine learning algorithms to classify cancer cell detection in its early stages.

2. Literature Survey

Classification in machine learning is established by making the machine learn a training dataset to store data. This learning can be classified into three types: supervised, semi-supervised, and unsupervised learning. In the supervised learning class, labeled data is present at the beginning. In semi-supervised learning, some of the class labels are familiar. Whereas, in unsupervised learning, class labels are not available. Once the training phase is finished, features are extracted from the data based on term frequency and then the classification technique is applied

Hajela et. al. [1] trained the classifier to teach the machine to detect cancer cells using features that were first used to detect cancer cells in the early stages. They have used a few algorithms to solve this problem, as the Convolutional Neural Network (CNN), image analysis, and the K-Nearest Neighbors algorithm. It has 91% specificity and 90% average accuracy. Image analysis scans the image and takes its fundamentals to give the desired output. CNN is very major in the field of deep learning. CNN uses 1-D, 2-D, and multidimensional convolutional models. It uses the softmax function, which converts a vector of N to the probability distribution of N possible outcomes. It takes images as input and gives an output that is easier to understand without losing the key features. K-Nearest Neighbor falls under the supervised ML algorithm, which is resolved by regression and classification problems. It predicts the output by calculating the distance between nodes, but by doing this, the time required for all calculations increases. Among all these algorithms, CNN is the best method with the best accuracy and specificity.

J. Awatramani et. al. [2] is on supervised and unsupervised learning features are used to detect malignant cells. The algorithms they experimented are Random Forest Tree, SVM (support vector machine), K-SVM (kernel-support vector machine), K-Nearest Neighbor, and Decision Tree. It has 98% accuracy in Random Forest Tree, which is one of the best methods in this paper. The random forest method uses a couple of decision trees where a decision tree with the maximum votes is selected. It is a type of unsupervised mode of learning. A Decision Tree is a method in which we gather a dataset and then create sub-data parts and again minimize it. It gives the output when the tree evolves, and at the termination level, we disclose the result. Here we use linear SVM, which is a type of supervised learning, and all data is segregated into hyperplanes, and the distance is considered between these hyperplanes. For KNN, we use Euclidian distance to calculate the distance between points of data. The sorting of data is done by taking the closest data related to the arrival of data. In kernel SVM we use a polynomial set, which is better to determine the performance as many datasets are non-linear.

H. Sami et. al. [3] uses ML methods to diagnose cancer and detect cancer cells. Here we use accuracy and susceptibility. The different methods used for detection and diagnosis are Sparse Compact incremental learning machines, Gauss-Newton Representation CNN, and Gene Expression Learning. The Sparse Compact incremental learning machine works on microarray gene expression data, which makes it robust against diverse noise and outliers. Due to its compact nature, it can also do classification tasks. The Gauss-Newton Representation uses sparse representation with training sample representation. It is useful to recognize a pattern. CNN predicts the result by considering the majority of votes. Gene expression learning tends to make the machine learn about the various genes present so that it can determine whether a gene is normal or abnormal depending on the data it contains.

R. Mosayebi et. al. [4] is about the detection of cancer cells in blood vessels using various methods? The methods are mobile nanosensors and molecular communication anomaly detection. Mobile nanosensors predict the detection of cancer cells by counting the biomarkers that flow independently in the blood vessels. If cancer biomarkers are observed, then the MN immediately warns according to the reading at different points by using the summation method. Molecular communication anomaly detection is known for the communication between different nano-machines which carry on communication between themselves and determine if cancer cells are present.

M. Vijay et. al. [5] uses Histopathological images to determine cancer cells. The techniques used here are simple CNN, dilated CNN, and Chanel-wise separable CNN. The simple CNN uses three layers, which are as follows: three max-pooling layers, three fully connected layers, and one output layer. The dilated CNN takes less time and has higher training than usual CNN. It also does not have a pooling layer as it skips dimension-making on image pixels and produces output based on softmax and SVM classification layers. It is also faster and takes less time to compute than simple CNN.

Yillin Yan et. al. [6] talks about algorithms, techniques, and applications. The algorithms and techniques are CNN, RNN, RvNN, and DBN. RvNN uses a tree-like structure, which is tender for NLP. RNN is appropriate for sequential information and is tender for NLP and speech processing. CNN was originally used for image recognition but is also used for NLP, speech processing, and computer vision. DBN is an unsupervised learning technique that is used on a directed connection.

Cruz-Roa et. al. [7] Is on image processing, visual interpretation, and automation. The techniques used in this paper are unsupervised and will follow a series of steps to get the desired output. An Autoencoder is used to get the most similar output for the given input. Image representation is accomplished through convolution and pooling by mapping the image by a set of k feature maps, and the original size of the image is increased by k in the process. Detection of BCC via Softmax, which is used in logistic regression, is used to calculate the theta vector which tells whether the input image is cancerous. This is considered by a value produced between 0 and 1 by using the sigmoid-activation function.

Abien Fred et. al. [8] is to detect breast cancer and to undertake consideration of visual processes and use algorithms. The techniques used are SVM, softmax, linear regression, and GRU-SVM. SVM has the highest accuracy among all the different algorithms used, which is 89.28%. GRU-SVM took 2 minutes and 54 seconds to complete its training. Linear classifiers are used as datasets, which means they have used linear regression and SVM.

Mehedi Masud et. al. [9] Is about the training of a convolutional neural network on breast cancer to get efficient output. Transferring knowledge is more feasible than starting to learn from scratch, which means that the CNN model does not need to learn everything from scratch; instead, it can transfer knowledge to the desired machine. Resnet, Alexnet, etc. models are tested to see which gives the best accuracy. Resnet has an accuracy of 96.4%. It had 152 layers and also consisted of a residual layer, which is very important in copying the image to the next layer. Performance metrics considered are accuracy, specificity, recall, and sensitivity. CNN models provide

automatic detection of breast cancer via ultrasound images.

M. E. Gamil et. al. [10] is about breast cancer detection in the early stages using image processing. The first block is about the filtering block, which scans the ultrasound image but, due to noise, a speckle is created which blurs the image. Filtering and smoothing help to remove noise from the image. The single image is divided into multiple segments to extract accurate information from it. The malignant value is considered and then the output is passed accordingly.

Priyanka Shahane et. al. [11] has been referred to get an idea to write a survey paper on machine learning.

Priyanka Shahane et. al. [12] has been referred to get a brief idea on survey paper and how to write a survey paper on related topics.

3. Performance Parameters:

After the dataset is prepared, it's time to use the machine learning algorithms to come up with an acceptable model which will predict the cancer stages upon feeding similar knowledge. Generating a model here suggests that coaching a machine learning model by feeding it knowledge in order that it will acknowledge a particular target and predict its values by feeding it additional knowledge while not the target.

The algorithms tried, during this case, were the simple regression model, the provision regression model, and therefore the rainforest regressed model of these models were trained against numerous sets of parameters to envision that setting provides the very best accuracy, and therefore the final parameters were:

'Specimen Type', 'Sample Type', 'DNA Input', 'TMB (no synonymous)','Sex',' diagnosing Age', 'Tumor Purity', 'Treatment'

The dataset was split into a magnitude relation of 75:25, wherever seventy five p.c of the dataset was utilized in coaching the model, and twenty five p.c in testing the accuracy.

Finally, the rainforest regress or model gave the very best accuracy with a delta of 0.7 once predicting a cancer stage. Precision is calculated as,

Precision: True Positive / True Positive + False Negative

Where TP (True Positive) test result detects the condition when the condition is present. FP (False Positive) test result detects the condition when the condition is absent.

Recall (Sensitivity/ TP Rate) can be calculated using the following,

Recall: True Positive / True Positive + False Negative

Where the FN (False Negative) test result does not detect the condition when the condition is present.

Where True Negative test result does not detect the condition when the condition is absent.

FP Rate can be calculated as,

False Positive Rate: False Positive / False Positive + True Negative

Accuracy can be calculated as,

Accuracy: sum (absolute (Expected Output – Actual Output)) / 2

4. COMPARATIVE ANALYSIS

Sr. No	Paper	Best classification technique	Other techniques tested	Dataset	Performance	Year of Publish	Author
1	Deep Learning for Cancer Cell Detection and Segmentation [1]	Image analysis and Convolutional Neural Networks (CNN)	K-Nearest Neighbors Algorithm, Microcalcification	CAD datasets	91% specificity and 90% average accuracy	2019	P. Hajela, A. V. Pawar and S. Ahirrao
2	Early Stage Detection	Random Forest	SVM,K-SVM, KNN,	non-linear	98%	2020	J. Awatramani

	of Malignant Cells: A Step Towards Better Life [2]	Method	and Decision Tree	decision boundary	accuracy in Random Forest Tree		and N. Hasteer
3	Machine Learning approaches in Cancer detection and diagnosis [3]	CNN	Gauss-Newton representation, Gene expression learning, Sparse compact incremental learning machine (SCILM)	Mammographic mass datasets	Accuracy and susceptibility	2017	H. Sami, M. Sagheer, K. Riaz, M. Q. Mehmood and M. Zubair
4	Early Cancer Detection in Blood Vessels Using Mobile Nanosensors [4]	mobile nanosensors (MNSs)	Molecular communication, anomaly detection,	NA	Accuracy not mentioned	2018	R. Mosayebi, A. Ahmadzadeh, W. Wicke, V. Jamali, R. Schober and M. Nasiri-Kenari
5	Diagnosing Cancer Cells Using Histopathological Images with Deep Learning [5]	Channel wise separable with dilated CNN	Simple CNN, Dilated CNN Channel wise separable CNN	RNA-Seq	99.7 accuracy	2021	S. K. V.N. and M. Vijay
6	A Survey on Deep Learning: Algorithms, Techniques, and Applications [6]	Convolutional Neural Networks(CNN)	Recurrent Neural Network (RNN), RvNN, DBN, DBM	NA	No accurate accuracy	2018	Yillin Yan, S.S Iyengar,Shu-Yeng Ching
7	Architecture for Image Representation, Visual Interpretability, and Automated [7]	Basal Cell Carcinoma (BCC)	Multi neural networks	BCC dataset	89.4% in F-measure and 91.4% in balanced accuracy	2018	Cruz-Roa, A.A., Arevalo Ovalle, J.E., Madabhushi, A., González Osorio
8	On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset [8]	SVM	Linear Regression, Nearest Neighbor (NN) search, and Softmax Regression,	70% for the training phase, and 30% for the testing phase, linear classifier as dataset.	9.28% test accuracy	2018	*Abien Fred M. Agarap*
9	Pre-Trained Convolutional Neural Networks for Breast Cancer Detection Using Ultrasound Images [9]	VGG16	DenseNet, ResNet	CNN	Around 99% accuracy	2021	*Mehedi Masud, M. Shamim Hossain*
10	Fully automated CADx for early breast cancer detection using image processing and machine learning [10]	Automated CAD	Logistic regression, anisotropic diffusion	Image dataset (ultrasound)	95.3% accuracy	2018	M. E. Gamil, M. Mohamed Fouad, M. A. Abd El Ghany and K. Hoffinan

5. CONCLUSIONS

From this survey we can conclude that the problem of detecting cancer cell in early stages can be solved by using various machine learning techniques/algorithms like CNN, SVM, RNN, Linear regression, BCC, KNN, Random Forest, and so on. Out of all these algorithms CNN and Linear regression have the best accuracy and can be used to retrieve an efficient and accurate output. It has 99.7% of accuracy. Further we can use deep learning and softmax regression to increase throughput.

REFERENCES

[1] P. Hajela, A. V. Pawar and S. Ahirrao, "Deep Learning for Cancer Cell Detection and Segmentation: A Survey," 2018 IEEE Punecon, 2018, pp. 1-6, doi: 10.1109/PUNECON.2018.8745420.

[2] J. Awatramani and N. Hasteer, "Early Stage Detection of Malignant Cells: A Step Towards Better Life," 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2019, pp. 262-267, doi: 10.1109/ICCCIS48478.2019.8974543.

[3] H. Sami, M. Sagheer, K. Riaz, M. Q. Mehmood and M. Zubair, "Machine Learning-Based Approaches For Breast Cancer Detection in Microwave Imaging," 2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), 2021, pp. 72-73, doi: 10.23919/USNC-URSI51813.2021.9703518.

[4] R. Mosayebi, A. Ahmadzadeh, W. Wicke, V. Jamali, R. Schober and M. Nasiri-Kenari, "Early Cancer Detection in Blood Vessels Using Mobile Nanosensors," in IEEE Transactions on NanoBioscience, vol. 18, no. 2, pp. 103-116, April 2019, doi: 10.1109/TNB.2018.2885463.

[5] S. K. V.N. and M. Vijay, "Diagnosing Cancer Cells Using Histopathological Images with Deep Learning," 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2021, pp. 148-152, doi: 10.1109/WiSPNET51692.2021.9419468.

[6] Yillin Yan, S.S Iyengar,Shu-Yeng Ching, "A Survey on Deep Learning: Algorithms, Techniques, and Applications", ACM 2018, Volume 51 , Issue 5

[7] Cruz-Roa, A.A., Arevalo Ovalle, J.E., Madabhushi, A., González Osorio, F.A. (2013). A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. MICCAI 2013. Lecture Notes in Computer Science, vol 8150. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40763-5_50

[8] Abien Fred M. Agarap. 2018. on breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (ICMLSC '18). Association for Computing Machinery, New York, NY, USA, 5–9. https://doi.org/10.1145/3184066.3184080

[9] Mehedi Masud, M. Shamim Hossain, Hesham Alhumyani, Sultan S. Alshamrani, Omar Cheikhrouhou, Saleh Ibrahim, Ghulam Muhammad, Amr E. Eldin Rashed, and B. B. Gupta. 2021. Pre-Trained Convolutional Neural Networks for Breast Cancer Detection Using Ultrasound Images. ACM Trans. Internet Technol. 21, 4, Article 85 (November 2021), 17 pages. https://doi.org/10.1145/3418355

[10] M. E. Gamil, M. Mohamed Fouad, M. A. Abd El Ghany and K. Hoffinan, "Fully automated CADx for early breast cancer detection using image processing and machine learning," 2018 30th International Conference on Microelectronics (ICM), 2018, pp. 108-111, doi: 10.1109/ICM.2018.8704097.

[11] Priyanka Shahane, Deipali Gore "A Survey on Classification Techniques to Determine Fake vs. Real Identities on Social Media Platforms,"IJRDT, 2018.

[12] Priyanka Shahane,"A Survey on Book Recommendation System" ,Volume 9, Issue 5, May - 2021