IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

A Review on Multiple Cancer Diagnosis using Machine Learning

Dr. Prof. Manisha Pise[1], Shreeya Gondhalekar[2], Mrunmayai Linge[3], Rohit Madderlawar[4], Karan Kalaskar[5]

Professor, Department of Computer Science & Engineering[1]
Students, Department of Computer Science & Engineering[2,3,4,5]
Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur, Maharashtra, India.

Abstract: *Cancer is a fatal illness that is typically triggered by the co-existence of several illnesses and genetic defects. Cancer cells are aberrant areas of the human body that are frequently fatal to specific body sections. It is essential to swiftly and correctly diagnose the condition in order to recognise what might be used to treat cancer at an early stage, frequently referred to as a tumour. Although the approaches taken to address these problems differ.The major reasons of mortality include convoluted histories, inaccurate diagnosis, and inadequate care. In this study, current advances in the detection of cancers using machine learning techniques for the breast, brain, and lung will be reviewed, evaluated, categorised, and discussed.*

The report concludes Cancer detection and therapy are aided by the ways in which machine learning approaches employ supervised, unmonitored, and deep learning. The outcomes of several cutting-edge methodologies are bundled with metrics for accuracy, sensitivity, specificity, and false positives. On benchmark data sets, metrics are contrasted. The challenges of prospective future employment are also covered.

Keywords: Artificial-Intelligence, Medical Image Analysis, Cancer diagnosis, Neural-network, Machine-Learning.

I. INTRODUCTION

"Being healthy" tops the Priority list of every human being. Healthy life is deserved by everyone. And most of the time our attitude depends on how we feel. Anything happening right or wrong! towards health, the only consideration over such situations is showing & maintaining "+ve Approach".

One such condition widely existing in our surroundings is CancerIt is a fatal condition typically brought on by the aggregation of several clinical changes and inherited ailments. Anywhere in the human body, cancerous cells can develop and cause deadly growths. Tumours that reside in specific areas, such as breast cancer, brain cancer, and lung cancer, can be identified and found using medical imaging tests.

Out of 18.1 million new cancer diagnoses, GLOBOCAN predicts that 9.6 million people died from cancer in 2018. Lung cancer is the greatest cause of mortality, accounting for 18.4% of all fatalities, followed by both melanoma and non-melanoma skin cancers (1.3%), colon cancer (5.8%), prostate cancer (3.8%), and breast cancer (6.6%). The most important and complex organ in the body is the brain.

Ineffective mitotic pathways regulate how morphological brain cells behave. Malignant cells have undergone morphological diversity as a result of this method, including changes to their strength and size. There are two primary divisions for brain tumours: low rates and high rates. A high-grade tumour grows violently and obstructs the brain's blood supply, while a low-grade tumour spreads slowly.

Numerous malignant brain cancers are also known as neuroepithelial tumours. These brain tumours make up 5% of cases, and patients often live for less than five years after diagnosis. Most cancer cells do not compare well to nearby cells. Glioblastoma is a typical type of tumour. Making a precise diagnosis of brain tumours is also essential. The primary technique for diagnosing brain tumours aids in tumour analysis from many perspectives and is painless. The main technique for finding brain tumours, for instance, is the examination of Magnetic Resonance Images (MRI).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

A lung nodule called a pulmonary nodule may be an indication of lung cancer. Pulmonary cancer must be detected early because it is a fatal disorder that shortens patients' lives. A benign or malignant nodule is something that forms in a circle. The cancerous nodule is growing swiftly, and similar nodules will quickly spread to the other organ. This is why it's so important to take care of malignant nodules as soon as they are found. A CT scan is most frequently used to diagnose lung cancer. More investigation is required if abnormal lung areas are to be identified after a CT scan.

The main cause of death for women is breast cancer. Significant cancer cases have been steadily rising for more than 50 years. Meenalochini et al. estimate that approximately 40.610 females in the US died of breast cancer in 2017. There are two types of abnormal cells in the breast: malignant and benign. Cancer cells are able to move to other organs and do more harm when they do so.Despite the fact that there aren't many malignant cells, both benign and large cells are produced. Fat and thick tissue have drawbacks, making it challenging to find malignant tumours early in the cycle. To categorise the breast tumour end, advanced automated or computerised technologies are typically required.

An ailment within the human body is far more significant in its place, especially if cancer has been diagnosed. We will now focus on two forms of malignancies to show how machine learning assists in the identification of these illnesses.

II. LITERATURE SURVEY

In 24 recent research publications, the computational methods for predicting breast cancer have been investigated. The summaries of each are given below. The models Chaurasia et al. developed can predict both benign and aggressive breast cancer. Utilised was a data set on breast cancer in Wisconsin. In the dataset, there were 699 incidences divided into two categories (malignant and benign), with 34.5 percent of them being malignant. The experiment was looked at using WEKA, a Waikato environment for knowledge analysis.

The prediction models were made using the three most popular data mining methods, Naive Bayes, RBF Network, and J48. The researchers used 10-fold cross-validation procedures to assess the impartial assessment of the three prediction models for performance comparison. Nine clinical features with integer values, including cell size regularity, are based on the effectiveness and accuracy of the techniques. The 16 occurrences with missing values were removed to produce the data set of 683 cases. 241 provided an assessment of the models' performance, whereas 458 (65.5%) were benign.

The Naive Bayes model fared the best, with a classification accuracy of 97.36%, according to the results of the trial. The J48 model placed third with a classification accuracy of 93.41%, followed by the RBF Network in second place with a classification accuracy of 96.77%. The researchers also ran sensitivity analysis and specificity analysis on the three algorithms to comprehend the relative contributions of the separate components to survival prediction. According to the sensitivity data, the prognosis component "Class" was by far the most important predictor.

Datasets

The predictor was created using the Wisconsin dataset and the LIDC/IDRI dataset. These data sets are frequently used to evaluate, compare, and test the effectiveness of cutting-edge cancer treatment methods. The source, training and test sets, and metrics for the efficiency of cancer detection, segmentation, and classification are also covered in this part. This dataset consists of 274 sets, 192 training cases (154 HGG and 38 LGG) from the University of Pennsylvania's Perelman School of Medicine, 82 trial cases, and 82 test cases.. The instructional images present superbly realistic situations with little fluff.

Five codes—1 for necrosis, 2 for edoema, 3 for tumor-free improvement, 4 for tumour enhancement, and 0 for everything else—are used to assess the veracity of the ground.

Brain Cancer and breast cancer

The primary regulator of the humanoid system is the human brain. A brain tumour develops when brain cells grow and divide abnormally, and brain cancer develops when brain tumours continue to grow. Computer vision is important in the field of human health because it eliminates the need for human judgement to produce correct results.MRI can identify minute items. Our paper focuses on the utilisation of several methods for finding brain cancer using brain MRI.

In a brain tumour, a four-degree cell arrangement is deformed. Grade 1 and 2 brain tumours appear to progress slowly, in contrast to malignancies in grades 3 and 4, which spread swiftly and are difficult to treat

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Grade 2 tumours are malignant and cancerous. In order to improve accuracy, the input picture is pre-processed to remove sound and non-brain tissue. The identification of malignancies involves a variety of crucial techniques. Organs other than the brain are removed using techniques referred to as BSEs (brain surface extractionThe simple non-local mean (FNLM), partial differential diffusion (PDDF), and Wiener filter are used to reduce noise and deplete contrast for enhanced contrast. The Otsu threshold, fuzzy Cmean, and k-mean approaches are the most widely used techniques for segmenting brain tumours. Similar to how CNN's popular concepts for segmenting brain tumours include Net architecture. To transform the fragmented images into mathematical explanations, hand-crafted features are acquired after segmentation.

At the moment, characteristics are segregated before being discovered using more exact methods. Form-based functions, Local Binary Patterns (LBP), Histogram Orientation Gradient (HOG), and Gabor Wavelet Transform (GWT) are the most often used extraction approaches. Among other filtering and reduction methods, the genetic algorithm (GA) and principal component analysis (PCA) are also used to improve function selection. Currently, CNN architecture is largely acknowledged as a helpful method for identifying brain tumours.

Numerous changes are made to enhance image transparency. Preprocessing methods including noise reduction, histogram equalisation, and tip enhancement promote improvements in optimum performance. In the class of the most recent model, weighting is used to solve the problem of class imbalance. The outcomes of each experiment are shown, and the validation details are examined by a learnt model using the same set of images. The accuracy score for ConvNet, an LSTM-based network, is 75%, and 80% of the overall fusion is 82.29%. The MICCAI Challenge database, which combines multi-modal brain tumour segmentation (BRATS) from 2015, 2016, and 2017, as well as the most current advancements in imaging technology, validates the aforementioned model.

These technologies are created by entropy, and they allow users to quickly and accurately recognise and interact with fused vectors to categorization units. In terms of the coefficient of dice similarity (DSC), the study found that it was 0.99 with 2015 BRATS, 1.00 with 2016 BRATS, and 0.99 with 2017 BRATS. However, they didn't use their fusion or any other classifiers to assess the effectiveness of their technique.the study found that it was 0.99 with 2015 BRATS, 1.00 with 2016 BRATS, and 0.99 with 2017 BRATS. However, they didn't use their fusion or any other classifiers to assess the effectiveness of their technique.

Lung Cancer

Lung disease is a significant global cause of lung death. In the case of a thin, diffuse tumour, many patients will benefit from surgical, percutaneous, and operational treatment. It is unfortunate that while few people exhibit no symptoms in the early stages of infection, 75% of lung cancer cases will not be discovered until they have spread to nodal and metastatic disease. Australian research found that the typical lung cancer survival rate was 15%. The LIDC/IDRI database has been used by several writers to record their work on the identification and classification of lung nodules.

Through the use of artificial intelligence tools and the analysis of CT data, machine learning techniques assist in the detection and assessment of lung nodules. These systems, which are also known as decision support systems, categorise, extract, segment, and preprocess pictures before studying them. The application of MCNN to capture nodular heterogeneity involves the extraction of distinguishing features from several levels. Nodular is captured by MCNN.

Using lung nodule scanning and annotation, the proposed LIDC-IDRI technique was assessed. Three CNNs in the MCNN paradigm create input parallel nodule patches of different sizes. The segmentation process has a 97% accuracy rate based on the LIDC database. Accurate pulmonary nodule detection depends on a number of patches from the lung picture produced by the Frangi filter. For a 4-channel neural network model that classifies nodules from four layers using the combination of two photo sets, radiologists' input is obtained. 80.06 percent of the sensitivity may be attained with 4.7 false positives per scan and a 15.1 false-positive sensitivity per sample.

They arrived to the conclusion that the patchbased learning technique increases efficiency across a range of classes and significantly lowers false positives in a huge amount of picture data. The stored autoencoder and SoftMax were used to conduct a thorough analysis of those false-positive findings. Collaborations between the largest database that is open to the public, the Image Database Network Initiative and the Lung Image Database, revealed that the suggested method significantly reduced false by 2.8 per scan with a chance for 95.6 percent sensitivity. a modern hybrid sensor for 3D

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

nodule processing that integrates the Active Contour Model (ACM) and, in the following stage, networks 3D neighbourhoods according to spatial features.

A hybrid functions vector (HOG-PCA) is produced by each candidate nodule's theoretical component analysis by combining its geometrical texture with its (PCA). Naive Bayesian, SVM, Adaboost, and two other grade models are used to complete the triage. Four new classes are added as a result of the abstraction process. The assessment (LIDC) is conducted using a sample from the Lungs Image Consortium.

III. CHALLENGES

The redesign of a cancer testing pipeline, comprehension of the cancer growth concept, development of preclinical models for the treatment of challenging tumours, early treatment, and creative techniques for organising and carrying out clinical research are the main obstacles to cancer detection and treatment. The disadvantage of deep learning systems is their requirement for enormous data sets, which may be inaccessible for cancer tissues. When there is a greater availability of gene expression data, this method may be more efficient and reveal more important trends. Machine learning techniques are very good at scaling too large data sets. Cancer prediction is still difficult due to the complexity and high dimensionality of these event.

IV. DATA & TECHNIQUES

The more generic phrases (large scale) data analysis and analytics, or artificial intelligence and machine learning, are often more suitable when discussing real operations. We employ the following machine learning methods in this study:

- **Convolution neural network(CNN):** One of the deep learning techniques that is particularly effective in processing and identifying pictures is convolutional neural networks (CNNs). Convolutional layers, pooling layers, and completely linked layers are only a few of the layers that make up this structure. Convolutional layers, where filters are used to extract information from the input picture such edges, textures, and forms, are the most crucial part of a CNN. The output of the convolutional layers is then sent after the feature maps have been down-sampled, the most crucial data has been saved while the spatial dimensions are reduced, and pooling layers have been employed. To forecast or categorise the image, one or more fully connected layers are then applied using the output of the pooling layers.
- **Decision tree algorithms:** Decision tree algorithms are a key component of effective machine learning classification techniques. They are methods of supervised learning that employ obtained and modified data to improve outcomes. Additionally, a variety of research, notably those pertaining to medical and health issues, commonly use decision tree algorithms for classification. Decision tree algorithms come in a variety of forms, such as ID3 and C4.5.But the most often used decision tree approach is J48. J48 is an extension of ID3 and an enhanced version of C4.5.
- **The K-nearest-neighbors (kNN)** algorithm is a straightforward approach for supervised learning in pattern recognition. Its ease of use and superior performance in the field of machine learning make it one of the most often used neighbourhood classifiers. The KNN method looks in the pattern space for the k training tuples with the greatest similarity to the unknown tuples. It then classifies new occurrences based on similarity metrics and keeps all precedent examples. The optimal number of neighbours (k), which varies from one data sample to the next, determines performance.
- **Support Vector Machine (SVM):** SVM is a supervised learning technique for categorising both linear and nonlinear data that was developed from statistical learning theory. By increasing the margin of hyperplane splitting, the SVM concurrently divides data into two classes over a hyperplane while avoiding over-fitting the data.
- **Negative Bayes:** It is a probabilistic classifier: The Bayes theorem is combined with strong (naive) independent assumptions in one of the finest classification algorithms. The value of the feature is independent of the value of any other features given the class variable, which is the conclusion that follows with the highest likelihood. It determines whether the provided tuple belongs to a specific class.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

- **The logistic model,** also known as the logit model, in statistics represents the likelihood that a certain class or event, such as pass/fail, win/lose, alive/dead, or healthy/ill, will occur. This might be broadened to encompass a range of things, such figuring out whether an image has a cat, dog, lion, etc. The sum would be 1 since each object in the image would have a probability between 0 and 1. Around the turn of the 20th century, logistic regression began to be used in the biological sciences. It was consequently applied in other social science fields. If the objective variable (dependent variable) is categorical, logistic regression is utilised.

Proposed solution for above discussion

From above discussion we come to know that it is sometimes difficult for the doctors, patients, and pathologists to communicate with one another while also giving them access to the necessary services as soon as possible .Early detection of the cancer is one of the ultimate goal which can help to reduce the death rate. So we need to develop a system which can help patients in detecting the cancer. The project's aim is to increase the model's overall accuracy to 95% while successfully predicting brain cancer, breast cancer, and lung cancer.The main focus should be on cancer patients and the real issues they encounter, starting with the checkup phase and continuing through the process of professional doctor consultation and, if necessary, therapies and treatments.

A new field of early-detection cancer research has been made possible by various computer-assisted cancer diagnosis and classification techniques, which have the potential to minimise manual system errors. The current study is divided into a significant number of sections that discuss the most recent methodologies, analyses, and comparisons for F-measurement, sensitivity, precision, and correct data sets for brain tumours, breast cancer detection, and lung cancer

REFERENCES

[1]. Jaiman, H. (2020). Survey on lung cancer detection using machine learning. International Journal for Research in Applied Science and Engineering Technology, 8(6), 1970-1974. doi:10.22214/ijraset.2020.6323

[2]. Cawley, J. C., Burns, G. F., & Hayhoe, F. G. (2012). undefined. Springer Science & Business Media.

[3]. Gross, L. (n.d.). undefined. Novartis Foundation Symposia, 76-104. doi:10.1002/9780470718902.ch9

[4]. Heidari, A. (2018). undefined. Theranostics of Respiratory & Skin Diseases, 1(1). doi:10.32474/trsd.2018.01.000102. The nature of leukaemia, cytology, cytochemistry and the morphological classification of acute leukaemia. (2017). Leukaemia Diagnosis, 1-68. doi:10.1002/9781119210511.ch1 undefined. (2011). Leukaemia

[5]. Diagnosis, 377-382. doi:10.1002/9781444318470.app1 Yin, T. (2015). Advanced Ultrawideband imaging algorithms for breast cancer detection.

[6]. Meenalochini, G., & Ramkumar, S. (2020). Survey of machine learning algorithms for breast cancer detection using mammogram images. Materials Today: Proceedings. doi:10.1016/j.matpr.2020.08.543

[7]. S., M. (2020). Survey paper on fraud detection in Medicare using machine learning. International Journal of Psychosocial Rehabilitation, 24(5), 4170-4174. doi:10.37200/ijpr/v24i5/pr2020130

[8]. Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. Journal of Infection and Public Health, 13(9), 1274-1289. doi:10.1016/j.jiph.2020.06.033

[9]. T, A. (2020). A survey on building an effective intrusion detection system (IDS) using machine learning techniques, challenges and datasets. International Journal for Research in Applied Science and Engineering Technology, 8(7), 1473-1478. doi:10.22214/ijraset.2020.30598

[10]. Cancer detection cancer detection - Google search. (n.d.). Retrieved from https://www.google.com/search?q=CANCER+DETECTIONCANCER+DETECTION&hl=enUS&sxsrf=ALeKk02nhVWUzSgMjMl2fIxuO5RBA3YPKw:1612966527390&source=lnms&tbm=isch&sa=X&ved=2ahUKEwi6-ZjHwN_uAhVwVBUIHS2eA4Q_AUoAXoECBMQAw&biw=1337&bih=586

[11]. https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[12]. https://towardsdatascience.com/buildinga-simple-machine-learning-model-onbreast-cancer-data-eca4b3b99fa3

[13]. Original data Set:http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic% 29

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

[14]. Confusion Matrix:https://tatwan.github.io/How-To-Plot-AConfusion-Matrix- In-Python/
[15]. https://seaborn.pydata.org/tutorial/axis_grids.html
[16]. https://seaborn.pydata.org/generated/seaborn.pairplot.html
[17]. https://seaborn.pydata.org/generated/seaborn.heatmap.html