WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Deep Learning Approach for Prediction of Breast Cancer

M. Sri Harsha [a], ***K. Pratima*** [b]

[a,b] *Department of CSE , GMR Institute of Technology , Rajam , AP-532127,India*

A B S T R A C T

Breast cancer is a prevalent and significant cause of cancer in women. Its incidence has increased recently, making it a common health issue. Early detection is crucial to managing the disease effectively. To spot the simplest way to manage breast cancer, classification is an effective approach. Here and now, deep Learning ways are being astronomically used in the breast cancer classification problem. The aim of the study is to achieve high accuracy and efficiency in classifying histopathological images to identify breast cancer using the Dense Net classifier. The study focuses on evaluating the effectiveness of each algorithm in terms of accuracy and efficiency. The Break His data-set is used for this purpose.

Keywords: Breast Cancer, DenseNet121, Histopathological Image classification, Prediction.

1. Introduction

The health-care industry commonly employs technology for storing and retrieving electronic medical records pertaining to patients, equipment, and other involved parties. In the diagnosis and management of hematologic diseases, cancer has always been a challenging diagnosis. Cancer is characterized by the rapid and abnormal development of cells, which is caused by a distinct combination of genetic and epigenetic abnormalities.Breast cancer predominantly affects women, with a lifetime risk. The chances of a positive outcome in breast cancer greatly rely on timely detection. Even if the symptoms may be moderate in the beginning, if treated quickly, the chances of survival greatly increase. The many approaches used for breast cancer screening methods comprise mammography, ultrasound-guided surgical biopsies, and fine needle aspiration cytology (FNAC). Mammography has a notably low cancer detection rate in cases where the patient has dense breast tissue, and 10% to 30% of cases go undetected.The research in question leveraged machine learning techniques for swift robot training and the development of predictive models that aid in decision-making. Through the examination of tumor size, machine learning can be used to detect and classify breast cancer in its early stages. Machine learning-based methodologies are particularly effective in addressing issues related to categorization and prediction. The utilization of ML techniques in breast cancer research has the potential to facilitate the identification and prediction of tumor presence or absence.

2. Related Works

The recent study on this type of models has proposed numerous results to the problem being addressed then. There were few works that have provided the solution to the breast cancer prediction.

The authors has proposed an model for predicting BC(breast cancer). According to the study, triadic negative bone tumors (TNBC) and non-TNBC can be classified with accuracy using machine learning (ML) algorithms. Among the machine learning methods tested, support vector machine (SVM) was found to be the most effective technique for accurately classifying BC into TNBC and non-TNBC categories due to its higher sensitivity, specificity, and reduced misclassification errors. When compared to the other four classification algorithms, SVM demonstrated the highest level of accuracy and best performance overall of 90%, a recall of 87%, and a specificity of 90%. The k-nearest neighbor (KNN) algorithm followed with an accuracy of 87%, a recall of 76%, and a specificity of 88%. [1]. The article evaluated the performance of supervised machine learning methods on the Wisconsin Breast Cancer data-set, including SVM, KNN, RF, ANN s', and supply regression. RF s and LR had the lowest accuracy of 95.7%, while ANN s' had the best accuracy of 96.57% among these techniques [2].

The essay explores six distinct supervised machine learning techniques, namely KNN, linear regression, decision trees, RF, SVM, and SVM with radial basis function kernel.Deep learning combined with ANN s' produced the highest accuracy, scoring 96.9%. SVM and random forest both scored at 95%, which was the second-highest accuracy. KNN and supply regression had the lowest accuracy results [3]. To address the current challenges in the automated diagnosis of BC - IDC (+) using whole slide images (WSI s') from the P Cam Kaggle data-set, the study proposes a hybrid deep learning model that combines CNN (Convolution Neural Network) and GRU (Gated Repeated Unit) architectures. This model employs multiple layers of CNN s' and GRU s' to effectively detect the presence of breast IDC (+) cancer. The proposed model is compared with CNN -BILSTM and CNN -LSTM models for BC -IDC (+) detection and achieved an accuracy of 86% using the described techniques [4].

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: author@institute.xxx

The SiGaAtCNN Input STACKED RF model was found to be more effective than other commonly used methods for predicting breast cancer prognosis. A comparative analysis was performed using the METABRIC data-set, which included 1980 cases, and the TCGA-BRCA data-set, which included 1080 cases, to compare the proposed method with existing techniques. [5]. This study's goal is to suggest a technique for locating and identifying breast cancers using ultrasound pictures. The classification of ultrasound images was performed using six distinct machine learning methods, including KNN, SVM, DT, and NB classification. The proposed approach had a 92% accuracy rate. [6]. A motorized system based on machine learning, the objective of the study was to enhance the accuracy of breast cancer diagnosis and minimize mortality rates by utilizing the Wisconsin Breast Cancer Data-set. To achieve this, the feature extraction technique of LDA was combined with the RF and SVM machine learning algorithms. Both RF and SVM classifiers were applied to the data-set, with RF achieving a 95.6% accuracy rate and SVM achieving an accuracy rate of 94.4% [7].

The primary objective of this study is to identify the most effective machine learning algorithms for predicting individualized survival rates of breast cancer patients and determining optimal treatment options.New numerical variable stratification techniques were also suggested. As therapies for instances with adjutant therapy and positive hormone receptors, the study first assessed the efficacy and outcomes of hormone therapy and chemotherapy. The alternative goal was to determine whether, in comparison to conventional methods, the semi-parametric Cox and non-parametric Kaplan-Meier stratification methods for the numerical variables in the data-sets offer a better understanding of the probability of mortality and recurrence. As a result, the machine learning algorithms' accuracy went up by 20 to 30% [8].

The research paper proposes two distinct noise reduction techniques that utilize Principal Component Analysis (PCA) for outlier removal and dimensionality reduction. Decision Trees, Logistic Regression, Bayesian approaches, Support Vector Machines, instance-based techniques, and Artificial Neural Networks are only a few of the machine learning algorithms to which the proposed methodologies have been used. The findings demonstrated that the PCA -based methods reduced attribute noise in the data-set successfully, obtaining an accuracy of 95% [9]. The article suggests a paradigm for automatically classifying breast cancer from histopathological pictures stained with H&E. (Hematoxylin and Eosin). Stain normalization, data augmentation, and ResNet-34, a deep learning-based classification system for identifying cancerous and benign samples, are all included in the framework. For the Break His data-set, the performance of the suggested algorithms was assessed for binary classification, with an average accuracy of 94.03%. In comparison to other methods, the suggested methods demonstrated an improvement in classification accuracy of 5% to 7% [10].

The proposed system was constructed using Apache Kafka and Apache Spark. To predict whether breast cancer was malignant or benign, LR, SVM, RF, and DT machine learning techniques were evaluated using features from the BCWD data-set. To assess these algorithms' effectiveness, an offline model was created. The outcomes demonstrated that RF with RFECV characteristics had the greatest accuracy of 96% [11]. This research aims to develop a computer-aided diagnosis (CAD) system that can assist radiologists in accurately identifying and categorizing breast cancer lesions in ultrasound images as either benign or malignant. The study made use of the Wisconsin Breast Cancer Data-set, and five classifiers—KNN, SVM, RF, XGBoost, and LightGBM machine learning algorithms to determine their effectiveness.The outcomes demonstrated that LightGBM performed better than the competition, obtaining an accuracy of 96.86% [12].

3. Materials and methods

The process of model structuring involves establishing methods for data collection, identifying, and prioritizing relevant information in the data to address specific questions, selecting a suitable statistical, analytic, or simulation model for gaining insights and making predictions. When building a machine learning model, the data is usually divided into two sets - the training set and the testing set - in a ratio of 80:20.

3.1 Data-set

For research purposes, the Break His data-set provides publicly available breast histopathology images. The data-set comprises 9,109 microscopic images of breast tissue samples obtained from 82 patients, collected from two distinct medical centers located in Brazil. The images are stained with Hematoxylin and Eosin (H&E) and are divided into two classes: benign and malignant.The data-set is organized into four sub - data-sets: benign tumors (2,480 images), malignant tumors (5,400 images), benign tissue (1,000 images), and normal tissue (229 images). The images have a resolution of 700x460 pixels and are in JPEG format.The Break His data-set is frequently employed to create and evaluate computer-aided diagnosis (CAD) systems that detect and categorize breast cancer.

3.2 Data Collection

Data collection is a crucial step in building any deep learning model. The performance and accuracy of a model are significantly impacted by the quality and quantity of data used for training.The process of collecting data for a deep learning model involves several steps, such as: Identifying the problem, Defining the scope, Data sources Data labeling Data cleaning ,Data augmentation Data splitting etc. Overall, data collection is a crucial step in building a deep learning model, and it requires careful planning and execution to ensure that the model is accurate and reliable.

3.3 Data pre-processing

Preparing data for deep learning models involves several essential steps, including data cleaning, normalization, feature selection, data augmentation, and encoding categorical variables. These pre-processing steps transform raw data into a format suitable for model training. In deep learning, data pre-

processing plays a crucial role as it can significantly impact the model's accuracy and performance. Thus, Histopathological images are re-sized in 256 * 256 pixels by calculating the pixels at each point checking maximum and min pixels.

3.4 Libraries

Some of the important libraries that can be used for the implementation of the proposed model are

Torch

PyTorch is an open-source machine learning framework built on Python and Torch library. It is widely used for building deep neural networks and is considered one of the most popular platforms for deep learning research. The framework is designed to facilitate the transition from research prototypes to implementation, and it supports almost 200 distinct mathematical operations. PyTorch's popularity continues to grow as it simplifies the process of building models for artificial neural networks. Data scientists mainly use PyTorch for research and AI applications.

Tqdm

Python Progress Bars are made using the tqdm module. The Arabic word taqaddum, which meaning "progress," is where it gets its name. One of the advantages of using progress bars is their ease of integration into loops, functions, and even Pandas. They not only provide an estimate of the execution time but also give an idea of the remaining time, which is particularly useful when working with large data-sets. Progress bars also provide visual feedback on the status of the code execution, allowing one to monitor the kernel's progress.

Pandas

A Python library for working with data sets is called Pandas. Pandas is a library that provides functionality for exploring, cleaning, analyzing, and manipulating data. It was created by Wes McKinney in 2008, and the name "Pandas" is a combination of "Panel Data" and "Python Data Analysis." Pandas is a powerful tool for analyzing large data-sets and drawing insights based on statistical principles. With Pandas, even messy and disorganized data-sets can be organized and made comprehensible. In data science, having access to relevant and accurate data is essential for making informed decisions and drawing accurate conclusions.

Scipy

SciPy is a library that extends the capabilities of NumPy. It is often abbreviated as Scientific Python and is commonly used in signal processing, statistics, and optimization. SciPy is an open-source library, and like NumPy, it is easy to use. Travis Olliphant, the creator of NumPy, is also responsible for the development of SciPy. Functions that are extensively used in NumPy and data science have been enhanced and added in SciPy. The majority of SciPy was developed in Python, while certain parts were also written in C.

3.5 Proposed Model

Densenet121 is a convolution neural network (CNN) architecture. It is part of the Dense Net family of CNN architectures, which are known for their efficient use of parameters and memory and their ability to achieve high accuracy on image classification tasks.The architecture of Densenet121 is based on the idea of dense connectivity, which involves connecting every layer to every other layer in a feed-forward fashion. This creates a very deep network with a large number of parameters, but the dense connections help to alleviate the vanishing gradient problem and promote feature reuse, which can improve training efficiency and model accuracy.Densenet121 consists of four dense blocks, each followed by a transition layer. The first dense block has 6 convolution layers with a growth rate of 32. The subsequent dense blocks have 12, 24, and 16 convolution layers, respectively, with growth rates of 32, 64, and 128.

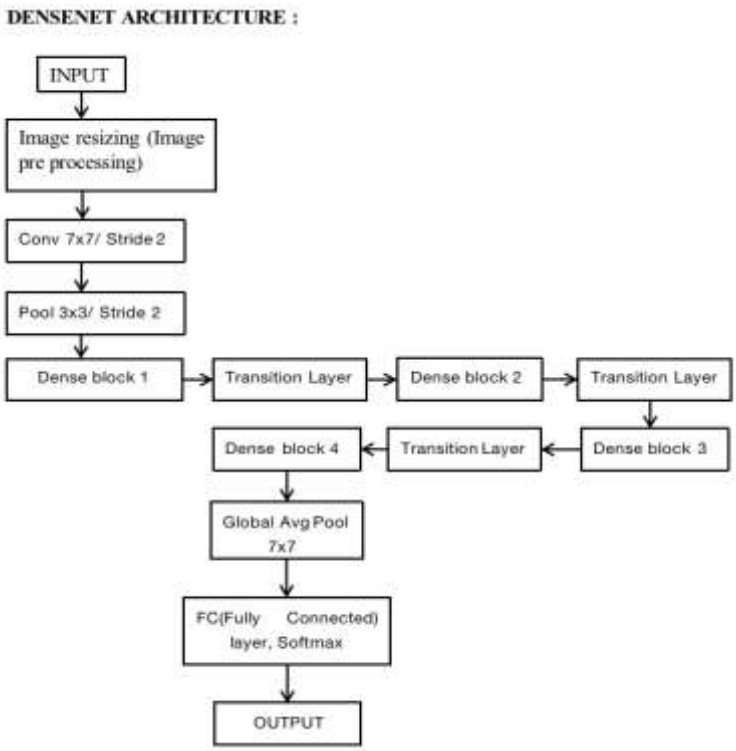

Fig. 1 - Model Architecture.

Prior to passing them on to the following dense block, Densenet121 uses transition layers to down sample the feature maps and lower the number of channels. They are made up of three layers: a 2x2 average pooling layer, a 1x1 convolution layer, and a batch normalization layer.The output feature maps are subjected to a global average pooling layer at the conclusion of the last dense block, which averages the feature maps' spatial dimensions to create a single vector. Next, the vector is fed into a fully connected layer containing 1000 units to produce class probabilities(equal to the number of Image-net classes).

Overall, Densenet121 is a powerful and efficient CNN model has demonstrated exceptional performance in image classification tasks, including the widely used Image-net benchmark data set. Figure 1 depicts the complete flow of dense-net 121 model including data pre-processing step

DenseNet-121

5+(6+12+24+16)*2=121

5- Convolution and Pooling Layer

3- Transition layers (6,12,24)

1- Classification layer (16)

2- Dense Block (1x1 and 3x3 conv)

In the convolution layers of Dense Net 121, a filter (also known as a kernel) is convoluted with the input image or feature map to produce a new feature map. This operation is repeated for multiple filters to produce multiple feature maps.

Figure 2 contains complete operations that performs in convolution layer which includes batch normalization and a rectified linear activation function (Rel U). Batch normalization helps to normalize the output of the convolution operation, making the network more stable and efficient. Rel U applies a nonlinear function to the output of the convolution layer, introducing non-linearity into the network.

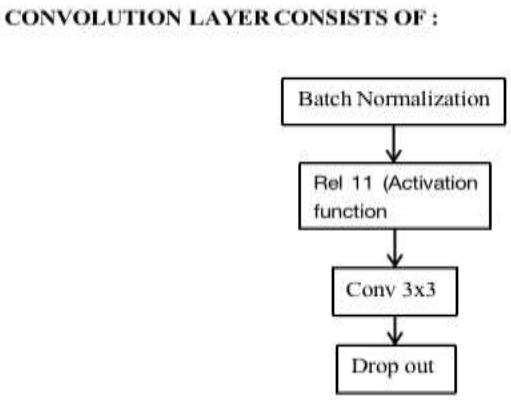

Fig. 2 - Convolution layer.

Between the dense blocks in dense net, three transition layers are employed to perform convolution and pooling. These layers lower the size of the feature map created by the pooling layer and are in charge of creating the activation map or feature map through the convolution layer. According to the study, the batch-norm layer, 1x1 convolution, and 2x2 average pooling layers make up the transition layers in the Dense Net design. The 1x1 conv basically down samples the data from n input features to n output features. Figure 3 contains the flow that goes under transition layer

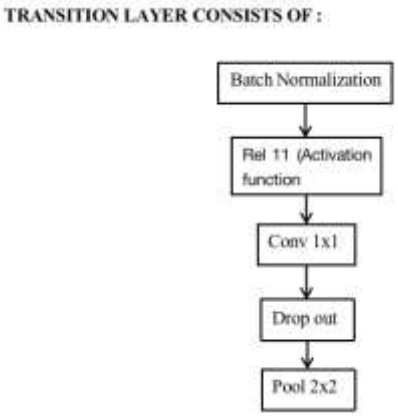

Fig. 3 - Transition layer.

3.6 Model Training

The _train_epoch method performs training for a single epoch, iterating over batches of data, computing the loss and gradients, and updating the model parameters using the optimizer. The method also updates metrics such as accuracy and logs the training progress to a logger object.

3.7 Model Testing

The _valid_epoch method performs validation after each training epoch, by iterating over the validation data, computing the loss and metrics, and logging the results Figure 4 depicts histopathological images which goes under 20% of data in data-set to perform validation / testing .

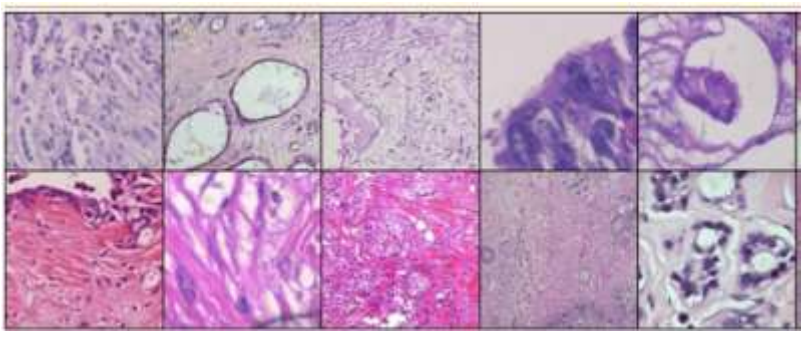

Fig. 4 - Histopathological Image.

4. Results and discussion

A deep learning model was developed using Visual Studio Code, which takes histopathological images as input and predicts the presence of breast cancer.

4.1 Experimental Results

Table1 demonstrated that our proposed model, Densenet121, achieved higher accuracy compared to other models such as SVM and Random Forest. Our model given comparatively greater results because it overcomes the disadvantages of SVM and Random Forest respective to size of data-set used and the features selected for classification of histopathological images.

Table 1 - Accuracy compared for breast cancer Prediction.

Algorithm	Accuracy Score
Random Forest	79
SVM	80.23
DenseNet 121	90.8

4.2 DenseNet 121 VS SVM

SVM employs more features than DenseNet, which uses less. As a result, it is possible to create a system that is better able to categorize histopathology pictures.

4.3 DenseNet 121 VS Random Forest

Densenet benefits from having big quantities of data and continually improving accuracy when it comes to massive data-sets. After a specific quantity of data is achieved, Random Forest frequently show no performance improvement.The results obtained are visualized by integrating tenser board to the directory. Visualization includes histograms, scalars and images. Results are obtained for both training and testing /validation sets. Below is a visualization of the accuracy of our proposed model on the training set.

Graph displaying the accuracy of proposed model on the validation set.

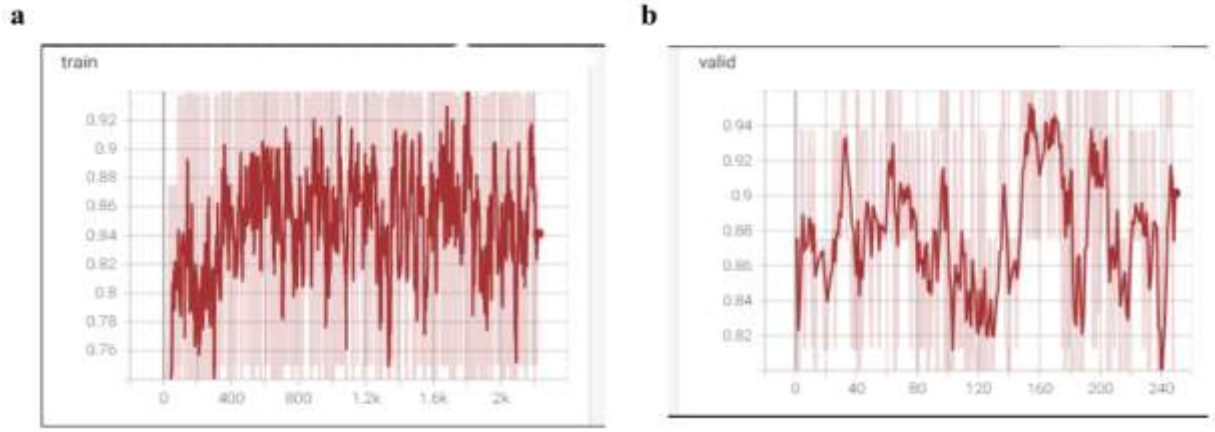

Fig. 5 - (a) Accuracy on Training set; (b) Accuracy on Testing set.

4.4 Criterion

Loss: Loss functions are utilized to evaluate in machine learning, the term "residual" refers to the disparity between the model's predicted output and the actual output. They provide a measure of how well the model is performing and are used to optimize the model parameters during training.

val_loss: The validation loss is determined by computing the disparity between the predicted results and the real results of the validation data-set. A lower validation loss indicates that the model is better at generalizing to new data, as it is making more accurate predictions on the validation set.

A graphical illustration of the loss of our proposed model on the training set.

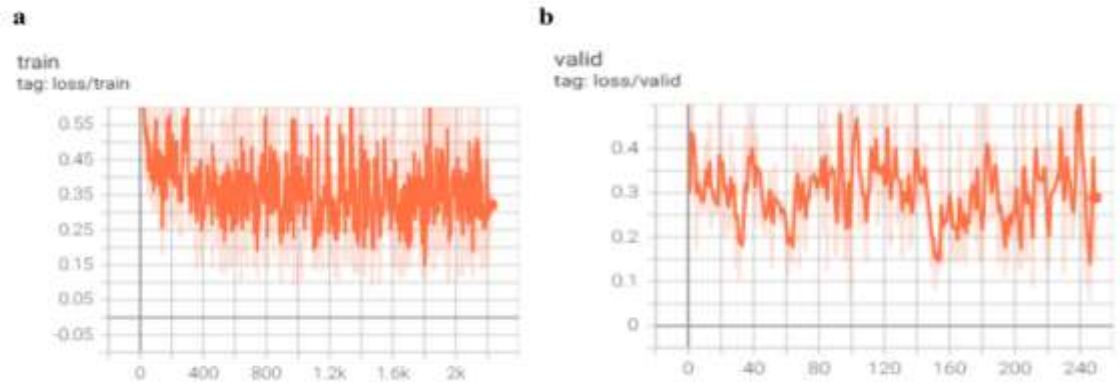

Fig. 6 - (a) Loss on Training set; (b) Loss on Validation set.

4.5 Metrics

Accuracy Score: Accuracy is a metric that measures the proportion of correctly identified cases from the total number of cases being assessed. Its highest possible value is 1, indicating perfect classification, while its lowest possible value is 0, indicating incorrect classification for all cases.

top_k_acc : This metric allows for multiple correct predictions and is useful in situations where the model may not be able to predict the exact label, but can still identify the correct category with some degree of confidence. The value of K is 2 and the result obtained of top_k_acc is 1.0, it means that the machine learning model accurately predicted the true class in either of the top 2 predicted classes for all the samples in the evaluation data-set. Obtaining a top_k_acc score of 1.0 is often the desired outcome when assessing the efficacy of a machine learning model, although this is an ideal situation.

val_accuracy : A val_accuracy score of 0.90 implies that the model can correctly classify 90% of the instances in the validation dataset. This is a substantial score, indicating that the model is performing admirably on the validation dataset.

val_top_k_acc: The result obtained by val_top_k_acc is 1.0, it means that the model accurately predicted the true class in the top-K most likely classes for all examples in the validation dataset. This scenario is considered ideal, as it suggests that the model can make accurate predictions with high confidence regarding the true class.

5. Conclusion and future scope

5.1 Conclusion

In this paper, after completion of implementation of the models like Densenet , SVM and Random Forest with data-sets . The performance of the proposed model in detecting breast cancer in histopathological images was found to be good. The integration with Tenser board helped to visualize the results with respective to accuracy and loss on training and validation sets .

5.2 Future scope

This paper only includes the accuracy and loss as metrics to evaluate results and for comparison of different algorithms implemented and only compared three algorithms named DenseNet 121 , SVM and Random Forest . Hence A survey analysis can be done to identify the best suited algorithm for detection of breast cancer in early stages.

Acknowledgments

It’s a great pleasure worked under the supervision of Dr. Aravind Karrothu , Assistant professor , Department of Computer Science and Engineering for his support and guidance throughout the project .

References

Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. Journal of personalized medicine, 11(2), 61.

Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1(5), 1-14.

Gupta, P., & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. Procedia Computer Science, 171, 593-601.

Wang, X., Ahmad, I., Javeed, D., Zaidi, S. A., Alotaibi, F. M., Ghoneim, M. E., ... & Eldin, E.T.(2022). Intelligent Hybrid Deep Learning Model for Breast Cancer Detection. Electronics, 11(17), 2767.

Arya, N., & Saha, S. (2021). Multi-modal advanced deep learning architectures for breast cancer survival prediction. Knowledge-Based Systems, 221, 106965.

Pourasad, Y., Zarouri, E., Salemizadeh Parizi, M., & Salih Mohammed, A. (2021). Presentation of novel architecture for diagnosis and identifying breast cancer location based on ultrasound images using machine learning. Diagnostics, 11(10), 1870.

Adebiyi, M. O., Arowolo, M. O., Mshelia, M. D., & Olugbara, O. O. (2022). A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. Applied Sciences, 12(22), 11455.

Maabreh, R. S. A., Alazzam, M. B., & AlGhamdi, A. S. (2021). Machine learning algorithms for prediction of survival curves in breast cancer patients. Applied Bionics and Biomechanics, 2021.

Ahuja, A., Al-Zogbi, L., & Krieger, A. (2021). Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification. Computers in Biology and Medicine, 135, 104576.

Magboo, V. P. C., & Magboo, M. S. A. (2021). Machine Learning Classifiers on Breast Cancer Recurrences. Procedia Computer Science, 192, 2742-2752.

Hu, C., Sun, X., Yuan, Z., & Wu, Y. (2021). Classification of breast cancer histopathological image with deep residual learning. International Journal of Imaging Systems and Technology, 31(3), 1583-1594.

Omran, N. F., Abd-el Ghany, S. F., Saleh, H., & Nabil, A. (2021). Breast cancer identification from patients' tweet streaming using machine learning solution on spark. Complexity, 2021.

Michael, Epimack, He Ma, Hong Li, and Shouliang Qi. "An optimized framework for breast cancer classification using machine learning." BioMed Research International 2022 (2022).

Arya, N., & Saha, S. (2021). Multi-modal advanced deep learning architectures for breast cancer survival prediction. Knowledge-Based Systems, 221, 106965.

Pourasad, Y., Zarouri, E., Salemizadeh Parizi, M., & Salih Mohammed, A. (2021). Presentation of novel architecture for diagnosis and identifying breast cancer location based on ultrasound images using machine learning. Diagnostics, 11(10), 1870.